The University Of Sheffield.

Statistical Modelling of Markers of Severity in Rheumatoid Arthritis

A thesis submitted to the University of Sheffield

to fulfil the requirement for the degree of

Doctor of Philosophy

Lyndsey Helen Taylor

Student Registration Number: 100104709

School of Health and Related Research

University of Sheffield

Lead Supervisor: M.D. Teare

Co-supervisor: A.G. Wilson

05 March 2014

Abstract

INTRODUCTION: Rheumatoid arthritis (RA) is a complex, chronic, autoimmune disorder. The severity of RA varies considerably between patients. Stratifying patients using effective prognostic biomarkers may facilitate therapeutic targeting of biological agents to those at risk of severe joint damage. Genetic variants including single nucleotide polymorphisms (SNPs) and environmental factors are known to associate with RA severity. However to date, no one has attempted to build a full predictive model for RA severity from these associated factors due to the high dimensional, highly correlated nature of the variables.

METHODS: Available data from a case-controlled study investigating genotype-phenotype associations in RA was used to investigate the predictors of RA severity (cases only). Using a sparse form of partial least squares (PLS) methodology, genetic SNPs and environmental factors were investigated to form a prediction model of a quantitative validated measure of erosive joint damage, called the Larsen score, before extending the methods to multiple RA severity measures. PLS is a dimension reduction technique which reduces the original variables to a linear combination with the influence of each variable being represented by a 'loading'. As 'loadings' are used to assess variable importance rather than beta coefficients from a regression model, PLS is not restricted by standard regression assumptions. Two sets of data were investigated; a genome wide association study (GWAS) recorded on 394 subjects referred to as 'GWAS SNPs' dataset and a maximum of 1009 subjects with 368 SNPs referred to as 'all subjects' dataset. A new method was developed to prevent over fitting of the PLS models which involved a three stage procedure. The first stage determined the order of predictive importance for the variables using 10 runs of 5, 7 or 10-fold cross validation (CV) (depending on the sample size). Absolute PLS loadings for each variable were ranked and the median calculated across the folds and runs to order the variables. The 'GWAS SNPs' dataset was analysed in 40 separate blocks of data. Variables ranked <200 were carried forward to a higher level model. The second stage investigated the number of variables to retain in the final model using an independent training and test set. The third stage tested the chosen model on a further independent set.

RESULTS: 'GWAS SNPs' dataset: Over fitted models containing 100 variables predicted well during CV (r=0.890). However, they performed poorly when tested on an independent set (r=0.385). Adding a second stage to the modelling prevented the over fitting, however only three variables were selected for the final model (disease duration, symptom duration and age at time of diagnosis) to achieve the highest correlation (r=0.622). '**All subjects' dataset:** Applying a three stage process resulted in a 10 variable model (disease duration, symptom duration, age at onset of symptoms, age at time of diagnosis, anti-citrullinated protein antibody (ACPA) category, ACPA value, body mass index (BMI), rs26510, DRB1 S2 and rs26232). The model predicted 182 independent subjects with a correlation of r=0.456. Analysing ACPA positive patients only increased the predictive correlation on an independent set (r=0.629), using a model with six variables (disease duration, symptom duration, age at onset of symptoms, age at time of diagnosis, BMI and rs2073839). Multiple Y variable modelling did not increase the ability to predict the Larsen score and other disease severity variables were poorly predicted.

CONCLUSION: SPLS is able to select key predictors of RA severity from a large dataset. A three stage approach is recommended to avoid over fitting of the model. Further research is required to investigate the success of the methodology of a more homogenous cohort.

Acknowledgements

Firstly I would like to thank my supervisor Dr Dawn Teare. When I entered your office ranting about how lost and confused I was, you were always so calm. Thank you for being such a good listener and for steering me in the right directions. I always felt better after our meetings and I can't thank you enough for your generous time and support. Particularly I also want to thank you for accepting me as a PhD student and giving me this opportunity to study. Thank you for providing me with such a superb research question, I have thoroughly enjoyed trying to answer it.

Thanks also to my co-supervisor Professor Gerry Wilson and the rest of the Department of Infection and Immunity at the University of Sheffield Medical School. You all made me feel so welcome. I learnt so much from your seminars even though I was continually embarrassed by claiming to be a statistician but never being able to match Dr Wilson's speed at mental arithmetic!

Many thanks to the Medical Research Council (MRC) for funding my PhD studies and for granting me additional funding through the MRC centenary award.

Thank you to the ScHARR medical statistics group. To Mike Campbell and Richard Jacques for helping me understand PLS theory. To Neil Shephard for the loan of books and programming tips and to Stephen Julious for the experience teaching, marking and general statistics (and life!) advice.

I want to give huge thanks to my husband Richard, for allowing me to take a break from industry to pursue research. Thank you for being so supportive and for always believing in me. Finally, I would like to thank my parents Anne and Bob, you gave me such a great start in life. You encouraged my love for mathematics/statistics and taught me that I can do anything I set my mind to.

Lyn Taylor 31st August 2013

Summary of thesis layout

The structure of the thesis generally reflects the order in which the work was completed. Motivation for the development of the methodology tended to come from the previous chapter's conclusions. The content of each chapter is summarised briefly below.

Chapter 1 provides an introduction to rheumatoid arthritis (RA) including the pathogenesis, aetiology, incidence, costs and current management. The chapter provides the medical justification behind why this research is useful and the overall aims of the project.

Chapter 2 introduces the data available for use on this project. The Genetics of RA (GoRA) study data contains disease phenotypes, demographics, environmental data and genetic measurements which are described in this chapter. Literature reviews are performed to identify reproducible genetic variants and environmental factors contributing to the severity of RA.

Chapter 3 presents relevant statistical multivariate methods and justifies the decision to use PLS to analyse the non-normally distributed Larsen score. Whilst PLS provides a method to reduce the dimension of the data for interpretation, Sparse PLS (SPLS) incorporates simultaneous variable selection and dimension reduction. Software packages capable of performing PLS are examined and R packages able to perform SPLS are compared. The theory behind SPLS is detailed and any requirements for data preparation discussed. Other general considerations regarding PLS are also detailed in this chapter.

Chapter 4 performs initial modelling using the 'percentage fold' method and the 'all subjects' dataset containing 912 subjects, 368 SNPs and 19 environmental variables. Initial model creation strategies are discussed in this chapter including the use of CV, how to avoid over fitted models, how to choose the number of components or variables for the model and how to achieve a robust model through using multiple runs. This chapter describes the method of retaining variables for the final model if they are selected in 80% (8 out of 10) of the CV folds in all runs of the modelling. It investigates the imputation of missing data using two methods ('quick' and Non-linear Iterative Partial Least Squares [NIPALS]). It explores transformations of the Larsen score but concludes no improvement to the modelling process. A flow diagram of the initial model fitting process is presented before reporting the results from modelling the 'quick' and NIPALS imputed datasets. To attempt quantification of how much of the model prediction is attributable to genetics, variance partitioning methods are researched using two approaches. Methods of univariate modelling to support the multivariate findings are explored using a Zero Inflated Negative Binomial model (ZINB) for the 'all subjects' dataset and a Negative Binomial (NB) model for the 'GWAS SNPs' dataset.

Chapter 5 investigates the impact of various imputation methods on a small subset of SNPs from the 'GWAS SNPs' dataset. The use of 'quick' imputation, NIPALS, IMPUTE2 and PLINK are discussed. The reproducibility of PLS is explored. Changes to the 'percentage fold' modelling methods which will be required to model the 'GWAS SNPs' dataset are researched including the percentage of folds a variable has to be selected (2 out of 5 folds or 3 out of 5 folds) and the selection of the number of variables to extract.

Chapter 6 performs SPLS modelling on the 'GWAS SNPs' dataset. To enable model fitting, the SNPs are split into 40 blocks. The performance of the 'percentage fold' method is compared, carrying forward to the higher level model, variables selected in 2/5 and 5/5 folds of 50 runs. . Alternative modelling strategies are explored to reduce running time and the requirement for user intervention. A new model creation strategy entitled the 'average rank' method is defined. This takes the median rank of the PLS loadings for each variable across the fold and runs. Any variables with an average rank in the top 200 are carried through to the higher level modelling. Attempts to replicate the top 10 SNPs found to be predictive of the Larsen score using SPLS modelling are investigated univariately by the Leiden University Medical Center (LUMC) using data from the Leiden Early Arthritis Clinic (LEAC) and North American Rheumatoid Arthritis Consortium (NARAC). The overall SPLS model performance is assessed and compared to using univariate modelling of SNPs.

Chapter 7 explores SIMCA software and compares the model created with the final model from chapter 6. Advantages and disadvantages of using SIMCA compared to R are discussed in addition to introducing orthogonal PLS (OPLS) modelling.

Chapter 8 attempts various methods of validation of the previous models created, to question whether they are over estimating the likely prediction ability of the model on independent data. For the 'GWAS SNPs' dataset, both the Larsen score data alone and the Larsen score and environmental data together are randomly permuted 100 times. The full modelling process is applied to the two new sets of 100 permuted datasets. The predictive performances are compared to the real Larsen score model. The 'GWAS SNPs' dataset is split into 80% of the patients for a training set and 20% of the patients for an independent test set to investigate over fitting. A new method is developed to prevent over fitting, entitled the 'two stage average rank' method. This uses 80% of the data for a variable ordering training set and 20% of the data for a variable selection training set.

The 'two stage average rank' method is applied to the 'all subjects' dataset before extending the method to a 'three stage average rank' method. This uses a 40% variable ordering training set, a 40% variable selection training set and a 20% independent test sample. The 'three stage average rank' method is applied to various subgroups based on disease duration and ACPA status to examine performance of a less heterogeneous sample.

Chapter 9 extends the research methods to modelling multiple measures of RA severity. RA severity and activity measures are grouped according to a principal components analysis (PCA) and three groups of variables are modelled using SPLS. Modelling is performed using the 'all subjects' dataset split into an 80% training dataset and 20% test dataset. In addition, the 'two stage' and 'three stage average rank' methods developed in Chapter 8 are used to model the multiple measures of RA severity.

Chapter 10 concludes the research, drawing together the findings from the above chapters. Justification of the PLS approach is provided with details of the contribution this research has made to the current research. Areas for future research are discussed in addition to a review of the software used.

Publications

The following papers were produced entirely whilst studying towards this PhD.

Taylor, L. H., Twigg, S., Worthington, J., Emery, P., Morgan, A. W., Wilson, A. G., Teare, M. D., (2013). Meta-analysis of the Association of Smoking and PTPN22 R620W Genotype on Autoantibody Status and Radiological Erosions in Rheumatoid Arthritis. *J Rheumatol,* 40**,** 1048-53.

Patwardhan, M., Pradhan, V., **Taylor, L. H.,** Thakkar, V., Kharkar, V., Khopkar, U., Ghosh K., Gawkrodger, D. J., Teare, M. D., Weetman, A. P., Kemp, E. H. (2013). The angiotensin-converting enzyme gene insertion/deletion polymorphism in Indian patients with vitiligo: a case-control study and meta-analysis. *Br J Dermatol,* 168**,** 1195-1204.

The following poster abstracts were produced entirely whilst studying towards this PhD.

Taylor, L. H., Twigg, S., UKRAGG., Worthington, J., Emery, P., Morgan, A. W., Wilson, A. G., Teare, M. D. (2012). Meta-analysis of the association of smoking and PTPN22 R620W genotype on autoantibody status and radiological erosions in Rheumatoid Arthritis. *Ann Rheum Dis* Vol. 71 (Suppl3) (pp. 156).

Taylor, L. H., Wilson, A. G., Teare, M. D. (2012). A Sparse Partial Least Squares Multivariate Model to Predict Rheumatoid Arthritis Erosive Joint Damage by Selecting Key Variables from a Large Panel of SNPs and Environmental Factors. *Genet Epidemiol* Vol. 36 (pp. 761).

Maxwell, J. R., **Taylor, L. H.,** Pachecho, R. A., Lawrence, N., Duff, G. W., Teare, M. D., Wilson, A. G. (2012). Inverse Relation between the Tumour Necrosis Factor Promoter Methylation and Transcript Levels in Leukocytes from Patients with Rheumatoid Arthritis. *Arthritis Rheum* Vol. 64 (pp. S427).

Table of contents

List of tables

List of figures

Abbreviations

ACPA	Anti-Citrullinated Protein Antibody
ACR	American College of Rheumatology
ANOVA	Analysis of Variance
BMI	Body Mass Index
C5orf30	Chromosome 5 open reading frame 30
CARD8	Caspase recruitment domain family, member 8
CASP	Critical Appraisal Skills Programme
CCA	Canonical Correlation Analysis
CCR	Correlated Component Regression
CCR5	Chemokine receptor type 5 (C-C motif)
CDK6	Cyclin-dependent kinase-6
CNV	Copy Number Variant
COX-2	Cyclooxygenase 2
CRP	C-Reactive Protein
CV	Cross Validation
DAS(28)	Disease Activity Score (28 item)
DD&SD	Disease Duration and Symptom Duration
DIP	Distal interphalangeal joint
DMARD	Disease Modifying Anti-Rheumatic Drug (s)
ERAP-1	Endoplasmic Reticulum Aminopeptidase 1
ESR	Erythrocyte Sedimentation Rate
FCRL-3	Fc receptor-like protein 3
GoRA	Genetics of Rheumatoid Arthritis
GSK	GlaxoSmithKline
GSTM1	Glutathione S-transferase Mu 1
GSTT1	Glutathione S-transferase Theta 1
GWAS	Genome Wide Association Study
HLA	Human Leukocyte Antigen
HWE	Hardy-Weinberg Equilibrium
ICER	Incremental Cost Effectiveness Ratio
IL	Interleukin
LASSO	Least Absolute Shrinkage and Selection Operator
LD	Linkage Disequilibrium
LEAC	Leiden Early Arthritis Clinic
LR	Likelihood ratio
LUMC	Leiden University Medical Center
MAF	Minor Allele Frequency
M(HAQ)	(Modified) Health Assessment Questionnaire
MBL	Mannose-binding lectin
MCP	Metacarpophalangeal joint
MHC	Major Histocompatibility Complex
MMP	Matrix metalloproteinase
MRC	Medical Research Council
MTP	Metatarsophalangeal joint
NARAC	North American Rheumatoid Arthritis Consortium
NB	Negative Binomial
NHS	National Health System
NICE	National Institute for Health and Care Excellence

NIPALS	Non-linear Iterative Partial Least Squares
OPLS	Orthogonal PLS
OR	Odds Ratio
(P)VAS	(Pain) Visual Analog Scale
PADI4	Peptidyl arginine deiminase, type IV
PCA	Principal Components Analysis
PIP	Proximal Interphalangeal joint
PLS	Partial Least Squares
PRESS	Prediction Residual Sum of Squares
PTPN22	Protein tyrosine phosphatase, non-receptor type 22
QALY	Quality Adjusted Life Year
QC	Quality Control
R^2-CV	R squared calculated under cross validation
RA	Rheumatoid Arthritis
RASEV	Measure self rating RA on a five point scale
RF	Rheumatoid Factor
RMSEP	Square root of the mean squared error of prediction
SEM	Structural Equation Modelling
STD	Standard Deviation
SF-36	Short form-36 health survey
SJC	Swollen Joint Count
SNPs	Single Nucleotide Polymorphism(s)
SPLS	Sparse Partial Least Squares
SS	Sum of Squares
TGF-β	Transforming Growth Factor Beta
TJC	Tender Joint Count
TNF	Tumour Necrosis Factor
UK	United Kingdom
VIP	Variable Influence on Projection
VNTR	Variable Number of Tandem Repeats
ZINB	Zero Inflated Negative Binomial

1. Introduction and aims

1.1. Introduction to rheumatoid arthritis

To provide background to the disease area, this chapter details the pathogenesis, aetiology, incidence, costs and current management of rheumatoid arthritis (RA). The motivation behind the research is discussed and the overall aim of the research defined.

1.1.1. Pathogenesis of rheumatoid arthritis

Rheumatoid arthritis (RA) is a complex, chronic, autoimmune disorder which mainly affects synovial joints such as the small joints of the hands and feet. It is a systemic disease which can also affect other areas such as the heart, lungs and the eyes. (National Institute for Health and Care Excellence (NICE) (2009)).

Joints consist of cartilage and a lining synovial layer covered by a fibrous capsule. The synovial lining consists of fibroblast-like synoviocytes that produce synovial fluid (a lubricating and nourishing fluid containing a high concentration of hyaluronic acid) and macrophages. In a synovial joint affected by RA, the lining layer of the synovium has an increased number of fibroblast-like and macrophage-like synoviocytes, macrophages and several populations of T cells and B cells (immune and inflammatory cells) which lead to hypercellularity (an abnormally high number of cells) of the synovium and increased blood flow (Isaacs and Moreland, 2002).

The cycle of activity starts with the joint becoming inflamed which leads to an increase in leukocytes and lymphocytes which in turn leads to the release of pro-inflammatory cytokines. The release of pro-inflammatory cytokines is followed by a proliferation of fibroblasts forming a pannus (an abnormal tumour like layer of fibrovascular tissue). Under these conditions the joint becomes hypoxic (deprived of oxygen) and growth factors are released stimulating angiogenesis (the formation of new blood vessels from pre-existing vessels). This in turn leads to an influx of leukocytes and the cycle continues with a further release of pro-inflammatory cytokines. These pro-inflammatory cytokines also lead to an activation of osteoclasts and chondrocytes which in turn leads to joint destruction such as cartilage and bone damage. Once the joint is damaged it can lead to joint failure with resultant pain and disability that can require the need for joint replacement surgery.

The severity of RA varies considerably between patients. RA which is uncontrolled can cause irreparable joint damage which can lead to disability, reduced quality of life, cardiovascular problems and other co-morbidities. Some patients experience a mild non-permanent disease compared to others who are subjected to a destructive debilitating disease with persistent inflammation (Cornelis et al., 2010). It has been reported that within two years of onset, subjects may experience moderate disability and it is estimated that after 10 years of onset, 30% become severely disabled. Approximately one third of patients stop work because of the disease (NICE, 2010).

1.1.2. Aetiology of RA

Whilst RA is the most common inflammatory joint disease, the aetiology of the disease remains unclear (Isaacs and Moreland, 2002). Current research acknowledges that both genetic and environmental factors contribute to risk (Plantinga et al., 2010). As RA is multi-factorial (caused by many factors), increasing the number of factors a patient has for the disease, increases the liability. Once this liability is above a certain threshold then the disease becomes expressed. In this context, heritability is defined as the proportion of this liability which can be accounted for by genetic variation.

Twin studies investigating the heritability of RA susceptibility estimate it to be in the region of 55% to 66%. Van der Woude et al. (2009) reported heritability to be 66% [95% CI: 44%-75%] and MacGregor et al. (2000) reported two cohorts with heritability estimated as 55% [95% CI: 40%-65%] in a United Kingdom (UK) cohort and 65% [95% CI 50%-77%] in a Finnish cohort. Regarding environmental factors there is increasing evidence of a complex interaction between smoking, genetic factors and the development of autoantibodies such as anti-citrullinated protein antibody (ACPA) and Rheumatoid factor (RF) (Kallberg et al., 2007, Morgan et al., 2009).

Numerous studies have been undertaken to evaluate the genetic influence on RA severity with limited success in replication. Although individual genetic variants have been identified, these have not been used to form a multi-variable prediction model. Such a model would use multiple genetic variants and environmental factors to identify patients who are susceptible to the severest form of the disease.

1.1.3. Incidence and prevalence of RA

RA affects approximately 1% of the adult population in the European Union although this can vary based on ethnicity and geography (Scott et al., 2010, NICE, 2010, Hochberg, 1981). NICE (2009) published estimates of 400,000 patients with RA in the UK with approximately 12,000 new cases a year. More recent literature by NICE (2010) increased this estimate to 580,000 patients with RA in England and Wales alone. Studies by Symmons et al. (1994, 2002) support these estimates, in addition to providing a detailed breakdown of disease incidence by gender and age group. Symmons et al. conclude that women have an earlier onset of RA, in addition to being three times more likely to get the disease as men. These estimates are further supported by Isaacs and Moreland (2002) who also hypothesise a recent potential decrease in incidence, particularly in women, which could be attributable to a protective effect of the oral contraceptive pill. Such claims are currently unsubstantiated. Section 2.2.2.2 provides more details regarding the typical age at onset and how this compares to the cohort studied in this research.

1.1.4. Costs of RA

Pugner et al. (2000) performed a Medline literature review to quantify direct and indirect costs associated with RA. Costs to the National Health System (NHS) and other healthcare support are considered direct costs of RA, whereas indirect costs to the economy include the impact of early mortality and loss of productivity. Using studies from the United States of America, Sweden, Italy, Canada, Netherlands and the UK, Pugner et al. concluded that the average annual per patient total cost was over $15,000 in 1998 (approximately £9300). Based on the NICE (2010) estimate of 580,000 patients with RA, this equates to a cost of £5.39 billion a year for England and Wales.

Previous NICE (2009) guidance reports that with approximately one third of patients stopping work due to RA within two years of disease onset, the total costs in the UK (including indirect costs and work related disability costs), is estimated between £3.80 and £4.75 billion per year. Therefore despite the relatively low incidence of RA, the cost to the UK economy is substantial.

1.1.5. Current management of RA

The primary aim of RA management is to modify the disease process, with resultant moderation of radiological progression leading to joint preservation and reduction in pain (NICE, 2009, Kwoh et al., 2002). NICE (2009) continues to add that as radiological progression is closely correlated with progressive functional impairment, reducing this impairment would result in reduced burden of costs on the NHS.

Disease modifying anti-rheumatic drugs (DMARDs) are synthetically produced and reduce synovitis and systemic joint inflammation which improves the function of the joint. Approximately 70% of patients treated with DMARDs respond initially, 40%-60% have a sustained response (NICE, 2009). NICE currently recommend that all patients are treated with DMARDs as the first line of therapy. Methotrexate is the leading DMARD which can be combined with other drugs of its type (Scott et al., 2010). If the patient has a disease activity score (DAS28) of 5.1 or more, confirmed on two occasions one month apart, and has already undergone two trials of two DMARDS (including methotrexate), then they should progress to use of a biologically produced treatment (a biologic) such as an anti-tumour necrosis factor (TNF) agent.

The reason for this two stage process is due to the cost effectiveness of methotrexate and other DMARDs compared to biologics. Biologics such as anti-TNF therapies are expensive. NICE (2010) estimated the annual cost per patient for treatment with Adalimumab, Etanercept, Infliximab, Rituximab or Abatacept was between £6984 to £10171 depending on the drug, dosage and course. The estimated incremental cost-effectiveness ratio (ICER) varies for each biologic depending on the treatment strategy. Chen et al. (2006, table 50) provide a summary of the estimated ICERs per quality adjusted life year (QALY) using a base strategy of DMARDs with no TNF inhibitor. They conclude that if following the NICE guidance for early RA, using TNF therapy as a 3rd line strategy based on early RA data, the ICER for Adalimumab, Etanercept and Infliximab compared to a base strategy of DMARDs alone (no TNF inhibitors), is between £28,000 to £35,000 per QALY. However, if used as a 1st line strategy in early RA, the Adalimumab and Etanercept ICER is between 49,000 and 170,000 per QALY depending on whether it is used in conjunction with Methotrexate. For Infliximab the estimated ICER per QALY increases to £650,000. Hence NHS treatments with anti-TNF therapies are cost effective only if used as a third line therapy in early RA.

1.2. Motivation and aims of this research

Patients have widely varying severities of disease even when adjusted for disease duration (Figure 2.2). This implies that some patients undergo a much faster rate of progression. As described in section 1.1.5, all patients receive the same treatment regimen and only when their disease fails to be controlled, do they progress to the more aggressive treatments.

If at disease diagnosis, patients could be stratified into those at higher risk of developing the more severe form of the disease, then this high risk group could be targeted earlier with a more aggressive treatment regimen. It has been shown that early and targeted management of RA symptoms leads to a reduction in the rate of disease progression (Agarwal, 2011, Kyburz et al., 2011, Teh and Wong, 2011).

To address the above motivation, the aim of this project is to use previously collected genetic and environmental data (described in section 2.2), to form a severity prediction model. If successful it is envisaged that this model could be used at disease diagnosis to estimate a patient's future predicted severity. The patients at highest risk of severe disease according to the model could then receive a different targeted treatment regimen which would slow the progression of their disease and hence reduce the economic cost to society in the long term.

2. Background

2.1. Aims

The aim of this chapter is to:

- Give background on the data which is available for use in this project
- Investigate correlations between RA severity measures
- Review the genetic and environmental factors which have previously been identified to be predictive of RA severity

2.2. Genetics of RA (GoRA) study data

A case-controlled study was sponsored by GlaxoSmithKline (GSK) to investigate the Genetics of Rheumatoid Arthritis (GoRA) (Protocol number: RARHD2001/0010/00). The study investigated the genotype-phenotype associations in RA and completed in 2006. Cases were unrelated RA patients attending clinic with moderate to severe persistent RA. To be included in the study they must have met the 1987 American College of Rheumatology (ACR) criteria for RA (Arnett et al., 1988) as assessed by their medical records when they were originally diagnosed. They were required to have evidence of at least one hand or foot erosion in the last three years and disease duration of greater than three years (Section 2.2.2.2 reveals the variation in subject's disease duration).

After conclusion of the study, further genotyping of samples of whole blood white cells was performed. This large quantity of data on the phenotypes and genotypes of RA patients will be used as the main resource for the formulation of RA severity prediction models and internal validation for the project. As the purpose of this research project is to investigate severity, comparisons between cases and control subjects are not included in this thesis. These details have previously been published (Marinou et al., 2007, Thomson et al., 2007). To enable the analysis, data collected by different research groups had to be combined into one dataset. The merging of the four data sources below resulted in one dataset containing 1009 subjects, 337887 single nucleotide polymorphisms (SNPs) and 126 other variables. See section 2.2.3.2.1 for further details about data manipulation, data cleaning and quality control. Ethical approval for a secondary data analysis has been obtained from School of Health and Related Research at Sheffield University (Appendix A).

- Sheffield: The main data from the GoRA study was stored in 62 individual spread sheets. Previous work by Dr James Maxwell (University of Sheffield Medical School) had combined some of the data together into a Filemaker Pro 8.5 database and into a single spread sheet. These two data sources were merged together and any discrepancies corrected against original source documents (patient files). Any additional data thought to be important from the spread sheets was also merged in. Genotyping was completed using the Taqman genotyping technology by Ioanna Marinou (University of Sheffield Medical School). This resulted in a dataset containing 45 SNPs, measures of disease severity and activity, environmental and demographic data on 1009 patients.
- Manchester: The data consisted of 943 patients who had been genotyped on 404 SNPs. Genotyping was completed using the sequenom platform with quality control steps described by Thomson et al. (2007). These SNPs were selected due to their previous associations with autoimmune disease.

- Milan: This data was a genome wide association study (GWAS) using the Illumina 370 copy number variant (CNV) chip consisting of 336076 SNPs measured on 397 patients.
- GSK: Data consisted of 2302 SNPs located in the Major Histocompatibility Complex (MHC) region genotyped by Illumina (San Diego, CA) on 855 patients.

2.2.1. Disease phenotypes (Y variables)

2.2.1.1. Summary of disease phenotypes

The following are quantifiable measures of rheumatoid arthritis severity recorded in the GoRA project. C-Reactive protein (CRP), Erythrocyte sedimentation rate (ESR) and disease activity score (DAS) are measures of disease activity. Modified health assessment questionnaire (MHAQ), pain visual analogue scale (PVAS) and short form-36 health survey (SF-36) are patient health assessments. The Larsen score is a measure of erosive joint damage assessed by radiography. The Larsen score can be broken down into the separate hand and foot scores in addition to a summary measure of any erosions (yes or no). Swollen Joint Count (SJC) and Tender Joint Count (TJC) variables count the number of swollen and tender joints. The variable RASEV measures how the patient rates their arthritis today on a five point scale from very mild to very severe.

Each phenotype is described below alphabetically with summary statistics on the GoRA population. Of particular interest is that most measures display large variation between the mean and median indicating skewed data and a wide range of severities.

C-Reactive protein (CRP): CRP is a protein found in the blood, the levels of which rise in response to inflammation. The GoRA data is positively skewed (mean=16.79, standard deviation (STD)=19.41, median=10, minimum to maximum=2.5 to 217.5).

Erythrocyte sedimentation rate (ESR): The ESR is the rate at which red blood cells precipitate in a period of one hour which provides a non-specific measure of inflammation. The data is positively skewed (mean=21.26, STD=16.67, median=17.00, minimum to maximum=1 to 128).

Larsen score: The Larsen score is a composite score measuring bone and cartilage damage. Boini and Guillemin (2001) provide a review of modifications made to the score since it was originally published in 1975. The version used on the GoRA patients was the modification made by Larsen in 1995 (Larsen, 1995). It measures 32 joints of the hands and feet. Each joint is rated as having no damage (score of 0) up to severe damage (score of 5). The maximum score for the most severe disease is 160. The data is positively skewed with an inflated number of zeros (Figure 2.1) (mean=36.45, STD=34.99, median=26.00, minimum to maximum=0 to 160). 13.6% (137/1009) of patients have no joint damage (erosions) at study entry (Larsen score of 0). 45% of subjects with a Larsen score of 0 were ACPA positive which is further described in section 2.2.2.3.

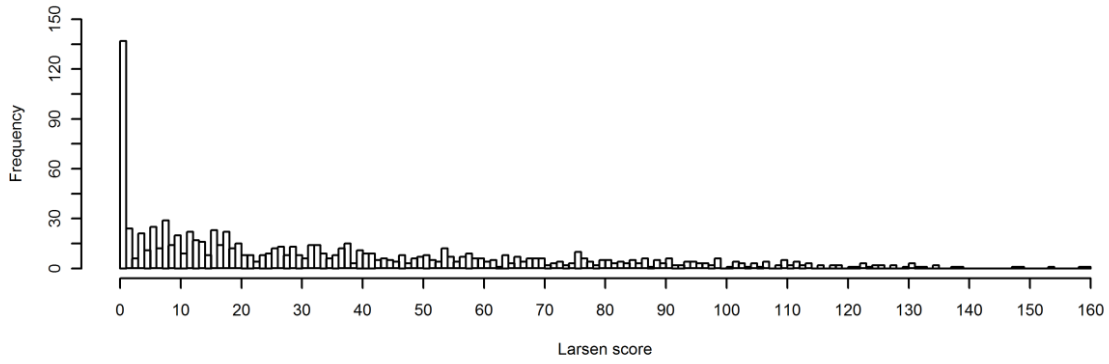


Figure 2.1 Larsen score distribution for GoRA subjects

Modified health assessment questionnaire (MHAQ): The MHAQ score consists of eight questions on health assessment. The four answers to each question are: Without any difficulty (0); With some difficulty (1); With much difficulty (2) and Unable to do so (3). The total score ranges from 0 to 24 with a mean =6.55, STD=4.88, median=6 and a range 0 to 24. It is approximately normally distributed except for an inflation in the number of subjects reporting a score of zero.

The MHAQ questions consist of:

1) Dress yourself including tying shoelaces and doing buttons
2) Get in and out of bed
3) Lift a full cup or glass to your mouth
4) Walk outdoors on flat ground
5) Wash and dry your entire body
6) Bend down to pick up clothing from the floor
7) Turn faucets / taps on and off
8) Get in and out of a car

Pain visual analogue scale (PVAS): To get a measure of pain, patients mark on a line (labelled 0 to 10), how they would rate their severity of rheumatoid arthritis pain which they are feeling today. The data is very slightly positively skewed as a higher number of patients seem to rate themselves with lower severity of pain (mean=37.04, STD=25.38, median=34, minimum=0 and maximum=100).

RA Severity (RASEV): Each patient was asked "How would you rate your arthritis today"? They selected from available responses of Very Mild, Mild, Moderate, Severe and Very Severe.

Short form-36 health survey (SF-36): SF-36 is a general health related quality of life metric. The data collected were recoded onto a 0-100 scale for each of the eight health concepts: physical functioning (a measure of physical activities including bathing or dressing), role physical (a measure of problems with work or other daily activities as a result of physical health), bodily pain (a measure of pain), general health (a measure of perceived personal heath), vitality (a measure of energy or fatigue), social functioning (a measure of interference with normal social activities due to physical or emotional problems), role emotional (a measure of problems with work or other daily activities due to emotional problems) and mental health (a measure of psychological distress and well-being) (Ware et al., 2000). The recoding provides normally distributed variables for analysis.

Table 2.1 reveals that on average GoRA patients score lower than the general population (<50) on the more physical concepts (bodily pain, physical functioning, role physical and vitality) and better than the general population (>50) on the more mental concepts (mental health, role emotional and social functioning).

Table 2.1 Summary of short form-36 health survey

Transformed SF-36 scores (N=1008)	Mean	STD	Minimum	Median	Maximum
Bodily pain	39	21	0	41	100
General health	43	20	0	42	100
Mental health	68	17	0	72	100
Physical functioning	37	28	0	30	100
Role emotional	64	44	0	100	100
Role physical	37	43	0	25	100
Social functioning	63	27	0	63	100
Vitality	39	19	0	40	100

Swollen joint count (SJC) and Tender joint count (TJC): The variables SJC28 and TJC28 assess whether the following 14 joints (on the left or right side) are swollen or tender respectively: shoulder, elbow, wrist, Metacarpophalangeal joint (MCP) I, MCP II, MCP III, MCP IV, MCP V, Proximal Interphalangeal joint (PIP) I, PIP II, PIP III, PIP IV, PIP V and knee. Both variables are positively skewed. For SJC28, mean = 5.75, STD=5.34, median=4, minimum=0 and maximum=28. For TJC28, mean =7.10, STD=5.88, median=5, minimum=0 and maximum=28. Extended measures, SJC and TJC, are also available which in addition to the above joints also include the following; Temporamandibular, Sternoclavicular, Acromioclavicular, Fingers (Distal interphalangeal joint [DIP] II, III, IV, V), Hip, Ankle, Toes (Tarsus IP I, II, III, IV, V), Metatarsophalangeal joint [MTP], MTP II, MTP III, MTP IV and MTP V.

Using the CRP, PVAS, TJC28 and SJC28 described above, the **disease activity score (DAS28)** is a composite measure calculated as:
$0.36*\log_e(CRP+1) + 0.014*PVAS + 0.56*sqrt(TJC28) + 0.28*sqrt(SJC28) + 0.96$
This is the leading European index, is simple to use and equally valid as the more comprehensive articular indices which can be time consuming in routine practice (Firestein et al., 2006). Due to the individual transformations applied to each of the composite parts, the overall DAS28 score is approximately normally distributed (mean =4.12, STD=1.22, median=4.06, minimum=0.96 and maximum=7.76).

2.2.1.2. Selection of primary and secondary disease phenotypes for analysis

The GoRA data is cross-sectional and patients are at different stages of their disease with different rates of disease activity. Measures of disease activity, inflammation and pain can vary considerably over time in accordance with how controlled the subject's disease is. Although erosive damage has been observed to repair itself (Keystone et al., 2004), cartilage damage is irreparable. Therefore, the Larsen score tends to mostly remain constant over time, or get worse if the disease is not controlled.

It was therefore decided to use the Larsen score as the primary phenotype and most important measure of RA severity to analyse in a single Y variable analysis. Figure 2.2 demonstrates the wide variation in the Larsen score which cannot be explained by the disease duration. This highlights the importance of finding other predictive markers of severity from the environmental and genetic data available.

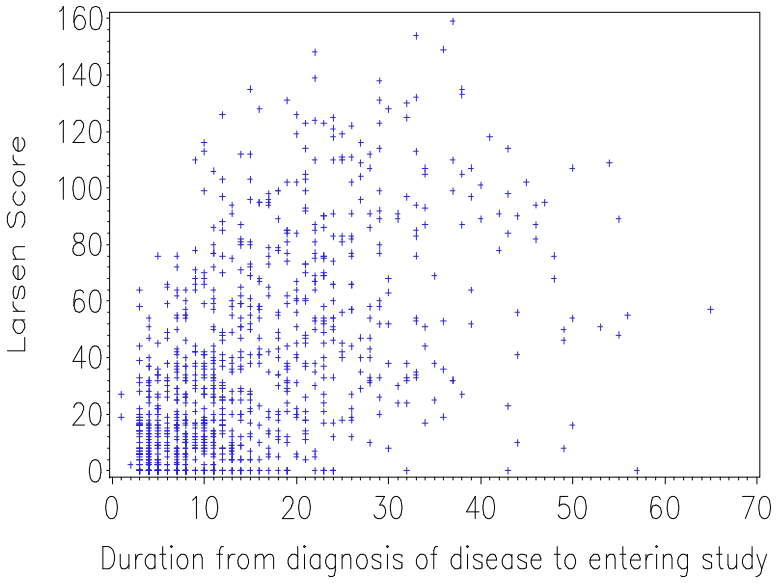


Figure 2.2 Larsen score plotted against disease duration for GoRA patients

Whilst the primary focus will be on the Larsen score, other phenotypes may also be important and may reveal different variables which are predictive of different types of RA severity such as pain, functional ability or laboratory measures of disease activity. Therefore, the other phenotypes will be explored using a multivariate Y variable analysis (Chapter 9). Garthwaite (1994), Frank and Friedman (1993) and Eriksson et al. (2006a, p. 99) suggest that only correlated Y's should be modelled together as they add stability to the model. If uncorrelated Y's are modelled together it tends to lead to a model which is difficult to interpret. Eriksson et al. (2006a, p. 99) suggest using a principal component analysis (PCA) of the Y-matrix to group the Y variables into suitably correlated groups. PCA is described in more detail in section 3.5.

In order to identify groups of phenotypes to be modelled together, PCA was performed in R Foundation for Statistical Computing, Vienna, Austria (version 2.13.1), using the package mixOmics (González et al., 2011, Lê Cao et al., 2009) version 3.0. Variables had mean centring performed and were scaled prior to extracting the first two components.

Figure 2.3 reveals correlations in the first and second principal components which represent 48.37% of the total variation. Ignoring the DAS28 variable, three groups can be identified situated in the different quadrants.

1) 8 domain SF-36 variables
2) 4 SJC/TJC variables
3) PVAS, RASEV, MHAQ, ESR, CRP, Any Erosions and the Larsen score (including the separate hand and foot counts)

The position of DAS28 on the graph reflects that it is a composite measure consisting of SJC, TJC, CRP & PVAS and is therefore situated between those two groups.

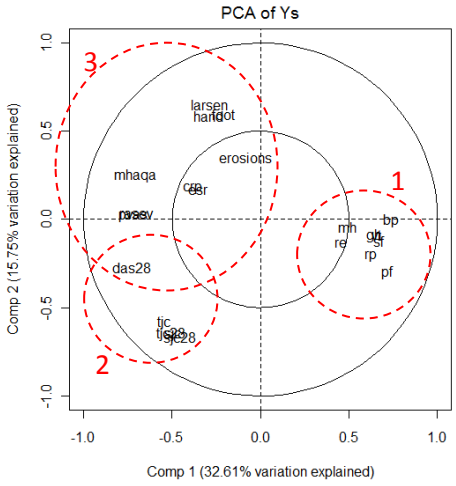


Figure 2.3 Loadings plot from PCA of RA severity measures (component 1 versus component 2)

For completeness, it was decided to model DAS28 with both groups 2 and 3 as it is correlated with variables in both groups. This work is shown in chapter 9.

2.2.2. Demographic, environmental and laboratory X variables

2.2.2.1. Summary of demographic characteristics

Table 2.2 and Table 2.3 summarise the key demographic characteristics. The mean age at entry to the study is 61 years with 72.7% being female, 20.1% currently smoking and 38.9% being former smokers. The mean smoking pack years is 13.04, however the data is positively skewed resulting in a median of just 3.5 pack years. Mean body mass index (BMI), height, waist circumference and weight are 25.7, 165 cm, 81cm and 70kg respectively. 68.4% drink alcohol, with 44.2% drinking more than five days a month, but only 10.5% having ever drunk more than five units almost every day.

Table 2.2 Summary of categorical demographics

		RA patients (N=1009)
Gender	Female	734 (72.7%)
	Male	275 (27.3%)
Smoking status	Smoker	203 (20.1%)
	Former smoker	393 (38.9%)
	Never	413 (40.9%)
Do you drink alcohol?	No	322 (31.9%)
	Yes	687 (68.1%)
During the past 30 days how many days did you have at least one drink of alcohol?	None	56 (8.0%)
	Less than 5	332 (47.7%)
	Between 5-10	143 (20.5%)
	More than 10	165 (23.7%)
Did you ever drink five or more drinks of alcoholic beverage almost every day?	No	903 (89.5%)
	Yes	106 (10.5%)

Table 2.3 Summary of continuous demographics

	N	Mean	STD	Minimum	Median	Maximum
Age at study entry (years)	1009	61.5	12.2	19.7	62.0	91.9
Body mass index (kg/m^2)	1009	25.7	4.8	12.5	24.9	50.4
Height (cm)	1009	164.9	9.5	137.2	162.6	200.0
Waist circumference (cm)	1007	81.0	12.0	50.8	81.0	194.0
Weight (kg)	1009	70.1	14.9	35.4	68.0	130.0
Smoking pack years	1009	13.0	18.3	0	3.5	150
Time since quitting smoking (Years) -Former smokers only	393	18.5	13.1	0	17	61
Average cigarettes per day	1009	9.2	12.4	0	5	200
Smoking duration (years)	1009	16.3	17.8	0	10	65

2.2.2.2. Summary of disease history

The mean (STD) age at time of disease diagnosis is 47.06 (15.0) and the mean (STD) age at disease onset is 44.18 (15.0) (Table 2.4). However, it is worth noting that there are patients included in this cohort who developed juvenile rheumatoid arthritis as they were diagnosed with RA when they were younger than 16 (N=42, five of whom were men). In addition, 386 patients (100 of whom were men) were diagnosed between the ages of 16 and 45. It is generally agreed that the incidence of RA increases with age (Scott et al., 2010). Although Isaacs and Moreland (2002) estimate that the peak age is in the 50s, NICE (2009) put this into the 70's for both genders. Symmons et al. (1994) (2002) report that male incidence of RA is rare under 45 years old. For women, they estimate incidence increases up to 45 years old, where it remains similar until 75 years old, before a fall in incidence in over 75s. As 38% (37% of males on the study) were diagnosed with RA before they were 45 years old, this cohort may not be truly representative of the population and may over represent patients diagnosed with RA at a younger age.

The mean (STD) duration from symptoms to entering the study is 17.30 (11.6) years and the mean (STD) duration since disease diagnosis is 14.70 (10.7) years. There is large variation in the duration patients have had the disease. This will be a key variable to consider in the analysis of disease severity. The duration of symptoms prior to diagnosis is skewed ranging from 0 to 41 years with a median of just one year and a mean of 2.54 years. If patients are recorded as being diagnosed before having symptoms, the duration of symptoms prior to diagnosis is set to equal zero, as it was assumed diagnosis could not be performed without any symptoms.

Table 2.4 Summary of age at and time since disease diagnosis and onset of symptoms

		Age at time of diagnosis	Age at onset of symptoms	Duration from diagnosis of disease to entering study	Duration from symptoms to entering study	Duration of symptoms prior to diagnosis
N	Valid	1007	1007	987	1005	986
	Missing	2	2	22	4	23
Mean		47.06	44.18	14.70	17.32	2.54
Median		47.90	45.20	11.00	14.00	1
STD		15.031	15.010	10.720	11.568	4.94
Minimum		4	2	1	3	0
Maximum		83	82	65	65	41

2.2.2.3. Summary of autoantibodies of RF and ACPA

The development of RA is preceded by and associated with elevated levels of autoantibodies in serum. ACPA epitopes have the highest specificity (proportion of ACPA negative subjects without RA) and sensitivity (proportion of ACPA positive subjects with RA) compared to other known autoantibodies associated with RA. Nell et al. (2005) estimate ACPA specificity at 98% and sensitivity at 41%. They also estimate high titre RF positivity (≥50 U/mL) to have 96% specificity and 45% sensitivity and compare this to using a lower threshold of >20 U/mL, which was found to have 89% specificity and 55% sensitivity. Bresnihan (2002) supports these claims reporting RF to have 91% specificity and 54% sensitivity. An association with more severe RA severity including radiographic damage when ACPA and RF antibodies are positive has been identified (Jansen et al., 2002, Ibn Yacoub et al., 2012a, Geng et al., 2012, van der Helm-van Mil and Huizinga, 2008).

In the GoRA study, a positive ACPA is defined as > 5.5 units/mL and a positive RF is defined as >40 IU/mL. Excluding missing data (112 patients), 62.0% of patients have positive ACPA and RF, 76.6% of patients have positive ACPA and 68.8% have positive RF (Table 2.5).

Table 2.5 Summary of RF by ACPA

	RF		
ACPA	Negative	Positive	Total
Negative	157 (17.1%)	57 (6.2%)	214
Positive	134 (14.6%)	568 (62.0%)	702 (76.6%)
Total	291	625 (68.2%)	916

Note: Percentages are out of the total number non-missing (916).

2.2.3. Genetic variant X variables

2.2.3.1. Summary of the shared epitope

The MHC region on chromosome 6 consists of molecules on the cell surface which regulate leukocytes (immune cells also called white blood cells) and their interactions with other body cells. The area (also called the human leukocyte antigen (HLA)) contains many genes spanning 3.6 mega bases. Marsh et al. (2010) provides a comprehensive list of HLA genes and alleles in this region. The list is periodically updated in accordance with latest research.

One gene of particular interest in this region is HLA-DRB1. A relationship was observed between amino acids in the third hypervariable region of the DR molecule and RA. Tezenas du Montcel et al. (2005) describe the work by Gregersen et al. (1987) who observed that the alleles reported to be associated with RA susceptibility all shared the R-A-A motif in positions 72-74 of the amino acid sequence. Patients with the R-A-A motif are said to have the 'Shared epitope'. Models including positions 67, 70, 71 and 76 have since been investigated attempting to quantify the risk of susceptibility and severity (Mackie et al., 2012a, Tezenas du Montcel et al., 2005, Michou et al., 2006, Meyer et al., 2011).

Tezenas du Montcel et al. (2005) group the DRB1 alleles at positions 70-74 into five categories: S1 (A-R-A-A or E-R-A-A), S2 (K-R-A-A), S3d (D-R-R-A-A), S3p (Q-R-R-A-A or R-R-R-A-A) and all other sequences to category X. As each subject has two alleles, Tezenas du Montcel et al. (2005) count the number of subjects with each distinct pair of alleles. For example, they count the number of subjects with both alleles S1 (S1/S1) and the number with one allele S1 and the other S3p (S1/S3p), until all combinations are accounted for. However, for predictive modelling, this would create many nominal categories, some with small numbers of subjects in. A reduced grouping method could be used (Michou et al., 2006), however this groups low risk categories together, rather than allowing them a different size of effect and could lose sensitivity of information. It was decided to follow the method used by Mewar et al. (2008) where the number of copies of each allele category are counted for each patient. This creates five variables, one for each category (S1, S2, S3d, S3p and X) and each subject has 0, 1 or 2 copies of that category. For example, a subject with S1/S1 would be said to have 2 copies of S1, 0 of S2, 0 of S3p, 0 of S3d and 0 of X. The GoRA data summarised using this method are shown in Table 2.6.

Table 2.6 Summary of shared epitope status

Category (N=1009)	Number of copies of each allele category		
	0	1	2
S1	632 (70.2%)	248 (27.6%)	20 (2.2%)
S2	449 (49.7%)	383 (42.4%)	72 (8.0%)
S3p	440 (48.8%)	401 (44.5%)	60 (6.7%)
S3d	829 (92.9%)	58 (6.5%)	5 (0.6%)
X	535 (59.1%)	303 (33.5%)	67 (7.4%)

To provide an overall summary, the number of patients with the shared epitope (defined as having the R-A-A motif) was also summarised. One hundred and seventy six GoRA patients (19.0%) have 0 copies of the shared epitope, 425 (45.8%) have one copy of the shared epitope allele and 326 (35.2%) have two copies of shared epitope alleles.

Many authors conclude the highest risk to susceptibility and severity is associated with the S2 category which has a positively charged lysine (K) encoding at position 71 (Tezenas du Montcel et al., 2005, Michou et al., 2006, Gyetvai et al., 2010, Meyer et al., 2011). However, results are more varied for the other categories. For example, there is some evidence that S1 and S3d are protective (or neutral) of severity (Michou et al., 2006, Tezenas du Montcel et al., 2005, Mackie et al., 2012a) and some evidence that all categories increase the risk of ACPA positive disease which in turn is associated with worse severity (Meyer et al., 2011, Gyetvai et al., 2010)

Many early studies attempting to use the HLA-DRB1 region to model RA susceptibility concluded that the role of this shared epitope could not fully explain HLA-DRB1 involvement (Tezenas du Montcel et al., 2000, Meyer et al., 1996, Genin et al., 1998, Rigby et al., 1998). More recently, van der Helm-van Mil and Huizinga (2008) and Mackie et al. (2012a) support this, suggesting a more complex relationship between HLA-DRB1 and RA. They suggest the effect is dependent on whether the subject is ACPA positive or negative which infers two separate distinct subgroups of disease. ACPA positivity has been found to be strongly associated with the shared epitope alleles although the association for RF with the shared epitope is thought to be weaker (van Gaalen et al., 2004, Irigoyen et al., 2005).

2.2.3.2. Single nucleotide polymorphisms (SNPs)

As described in section 2.2, SNP data measured on the GoRA subjects is obtained from four sources. Unfortunately, each data source measured a different number of GoRA subjects and genotyped different SNPs, sometimes with overlap.

Table 2.7 summarises for each data source, how many of the GoRA subjects' samples were attempted to be genotyped and the number of SNPs measured. It reveals that although there is a lot of data available, not all subjects have all SNPs recorded. A brief description of how the SNPs were selected by each data source and how they were genotyped can be found in section 2.2.

Table 2.7 Data collected by erosion status

Data source	Number of GoRA subjects[1]	Number of SNPs[2]	Number of usable subjects[3]	Number of subjects with erosions	Number of subjects without erosions
Sheffield	1009	45	1008	871	137
Manchester	943	404	943	834	109
GSK	855	2302	854	853	1
Milan (GWAS)	397	336076	394	386	9

[1] Number of unique subject identifiers in the data source.

[2] Number of SNPs available for use (see 2.2.3.2.1 for more details)

[3] Subjects were merged by a unique subject identifier. In a few cases, this was not possible (perhaps due to typographical errors) resulting in subjects with completely missing RA severity or environmental data or both. Hence these subjects were excluded from the analysis.

2.2.3.2.1. SNP quality control

PLINK (http://pngu.mgh.harvard.edu/purcell/plink/) (Purcell et al., 2007) was used to evaluate the distribution of missing data and to test for Hardy-Weinberg equilibrium (HWE) in the Milan GWAS data. Table 2.8 reveals subjects to have an acceptable subject call rate with a maximum for any subject of 5.8% of SNPs not able to be genotyped (median=0.6%). However, the SNP call rate identifies more of an issue as some SNPs have 100% data missing. On further inspection, 969 SNPs have >50% of subjects with missing genotyping (3720 SNPs with >15% of the subjects with missing genotyping). Traditional GWAS quality control (QC) techniques are likely to have excluded these SNPs at this stage of the research, however, as it was initially believed SNP QC to have already been completed on this data, these SNPs were not excluded until during the modelling described in section 4.2.3. Figure 2.4 and Figure 2.5 present the distributions of missing data for subjects and SNPs.

Table 2.8 Missing data and HWE p-values for Milan GWAS data

	n	n Missing	Mean	STD	Min	Median	Max
Subject % missing	397	0	0.9	0.8	0.3	0.6	5.8
SNP % missing	336076	0	0.9	5.7	0	0	100
HWE p-values	335189	887	0.556	0.317	1.5×10^{-42}	0.562	1

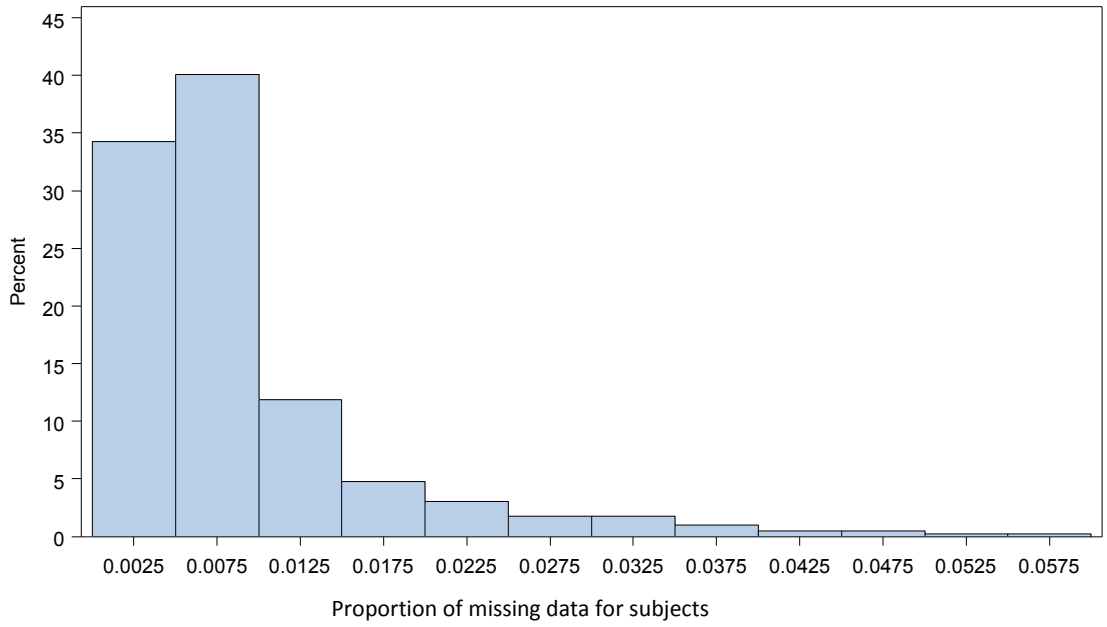


Figure 2.4 Distribution of Missing data for Subjects in the Milan GWAS data

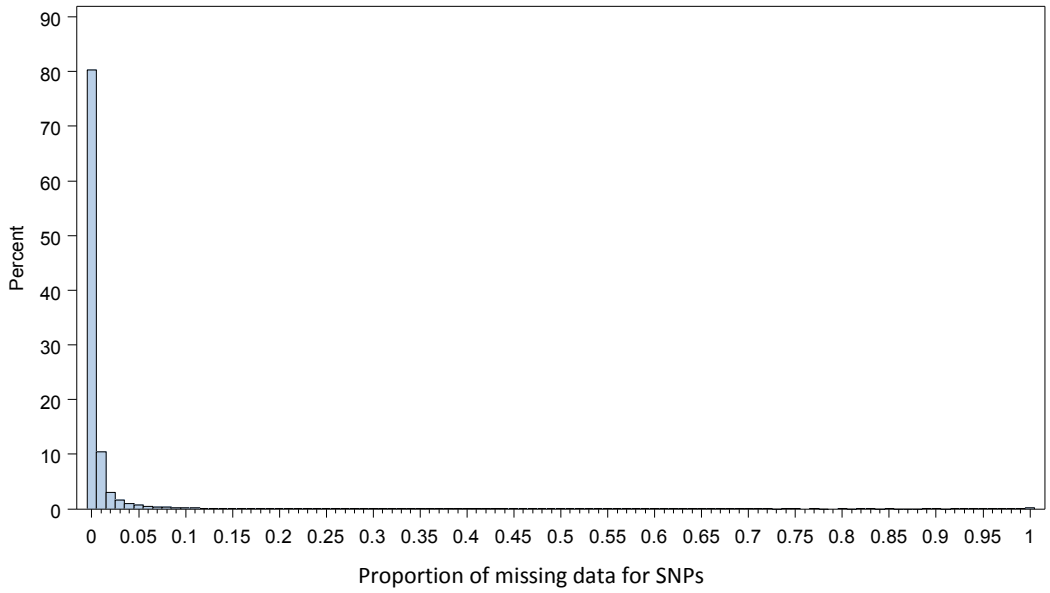


Figure 2.5 Distribution of Missing data for SNPs in the Milan GWAS data

Table 2.8 also presents a summary of the p-values from tests for HWE. The distribution of p-values is plotted in Figure 2.6. There is an increased number of p-values equal to 1 due to monomorphic SNPs being included in the data. The remaining distribution looks approximately uniform indicating no deviation from the expected distribution of p-values under the assumption of HWE for the GWAS SNP data.

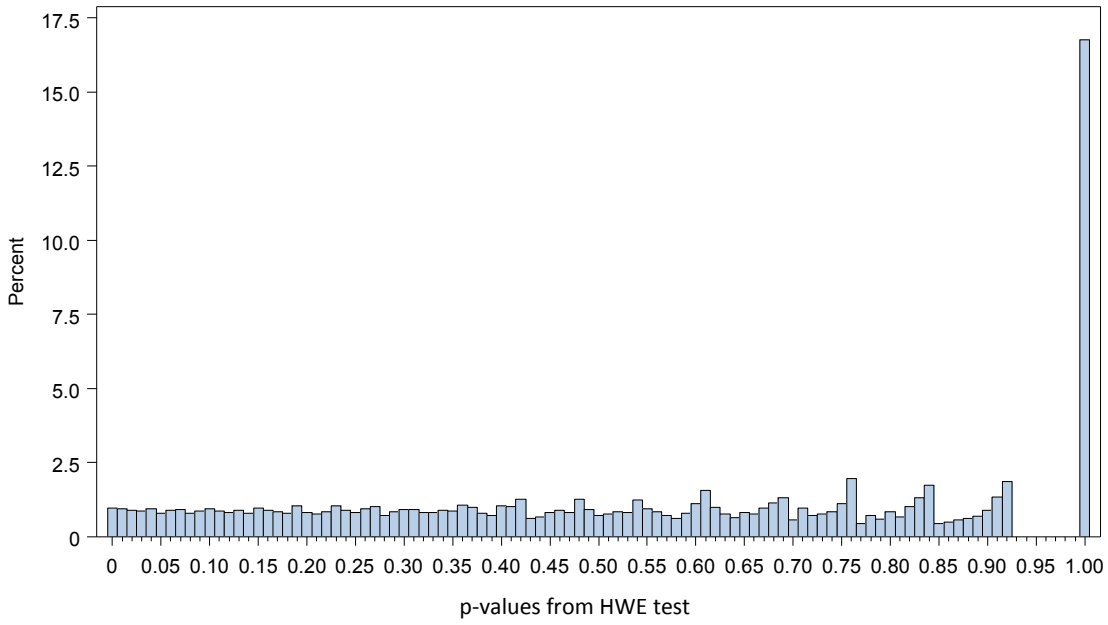


Figure 2.6 Distribution of p-values (Hardy-Weinberg equilibrium test) for Milan GWAS data

Data from the other sources was harder to check from a quality perspective due to:

- Genotyping in the Sheffield source was performed one SNP at a time and not all subjects' samples had been intended to be genotyped for all SNPs. For example, zero SNPs had <35% of the 1008 subjects genotyped, however 29 SNPs had between 35% and 50%. Zero SNPs had >50% and <85% and 17 SNPs had >85%. The peak in number of SNPs with between 35%-50% of SNPs genotyped indicates that it was never the intention to genotype all subjects for all SNPs.
- The Manchester and GSK data was extracted from large databases containing patients from other studies. Some SNPs had completely missing data for the GoRA subjects however it was likely that this was not due to failed genotyping but because the GoRA samples never intended to have that SNP genotyped.

It was therefore not possible to retrospectively decipher what the original call rates were for the Sheffield, Manchester and GSK data. Figure 2.7 presents histograms of the proportion of subjects with present data for each SNP, presented alongside the distribution of p-values from the HWE test. Although there were some SNPs with substantial missing data, it was not possible to tell if this was due to data quality or study design. Hence all SNPs were retained for the analysis and the issue of missing data was re-addressed in section 4.2.3.

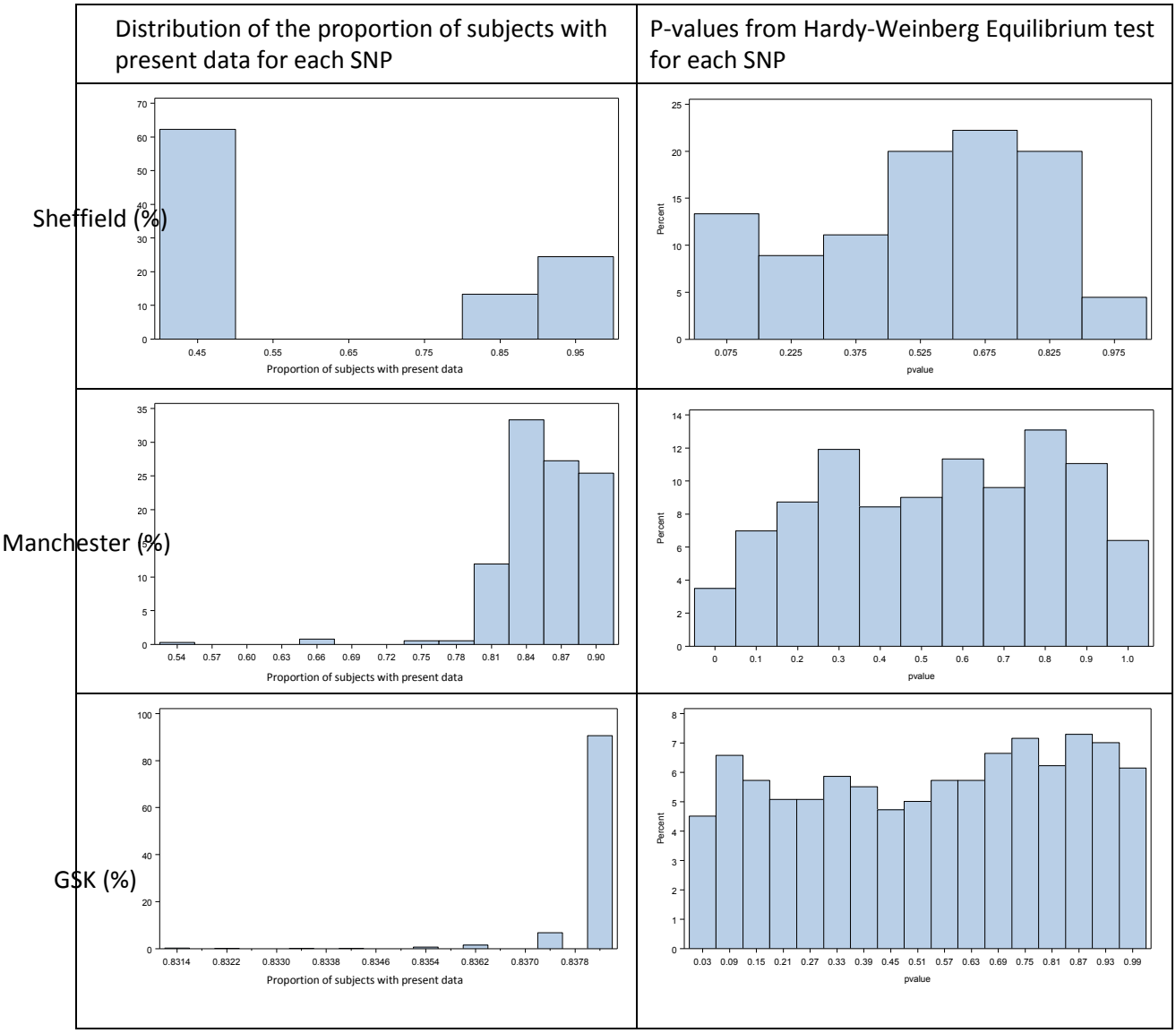


Figure 2.7 SNP quality checking for Sheffield, Manchester and GSK data

2.2.3.2.2. SNP data cleaning

All data manipulation and cleaning was performed in SAS/STAT software, Version 9.2 of the SAS System for Windows Copyright © 2002-2008 SAS Institute Inc. Cary, NC, USA. SNPs recorded by more than one source (Sheffield, Manchester, Milan or GSK) were compared. Any discrepancies were set to missing (blank) after checking the strand direction.

The number of SNPs compared and which source they came from, is presented in Table 2.9, along with summary statistics of the percentage of patients with discrepancies. Three SNPs (rs1295686, rs2516714 and rs2157337) had 26%, 27% and 17% disagreement across sources. rs1295686 was compared across the Sheffield & Milan data. rs2516714 and rs2157337 were compared across the GSK and Milan data. On the assumption that two reliable platforms were used for the genotyping, then this level of disagreement would indicate that this data is unreliable and should not be used in any analysis. However, after discussion regarding which genotyping method is the more reliable (for example would a GWAS be more reliable than using Taqman genotyping or visa-versa), it was

decided to set any disagreements across data sources to missing and providing that SNP still had sufficient data to be included in the analysis, then the SNP data was included. In hindsight, this may not have been the best approach, however, as only 930 SNPs were able to have this comparison performed out of the 337887 measured, it was considered unlikely that this would affect the results. After a RA severity prediction model has been created, the SNPs selected as predictive of RA severity could always be retrospectively checked to ensure the quality of the data was adequate.

Table 2.9 Percentage of patients with disagreement between two or more SNP data sources

Datasets compared (Number of SNPs compared)	Mean	STD	Min	Median	Max
Gora vs. Manchester (3 SNPs)	0.425	0.531	0.112	0.127	1.038
Gora vs. Milan (6 SNPs)	6.006	9.983	0.524	1.629	26.100
Gora vs. Milan (5 SNPs) –Exc. rs1295686*	1.656	1.849	0.524	1.050	5.128
Manchester vs. Milan (94 SNPs)	0.650	0.650	0.000	0.549	3.892
Gora vs. Milan vs. Manchester (2 SNPs)	1.150	0.660	0.683	1.150	1.616
GSK vs. Milan (819 SNPs)	1.572	1.307	0.000	1.575	27.297
GSK vs. Milan (817 SNPs) –Exc. rs2516714 and rs2157337*	1.517	0.772	0.000	1.575	7.710
Gora vs. Milan vs. GSK (6 SNPs)	1.222	0.689	0.655	1.027	2.559

*Exc. = SNPs excluded from the comparison.

2.2.3.2.3. Final SNP data for analysis

The Milan GWAS SNP data was only collected on 397 of the GoRA subjects which after merging with the other data sources resulted in only 394 subjects with usable data (due to unique subject identifiers not being consistent across sources). This resulted in approximately 60% of GoRA subjects having the majority of SNPs missing. It was therefore decided to create two datasets for use in the analysis to prevent having the problem of large quantities of missing data.

The first dataset hereafter called the 'all subjects' dataset, consists of SNPs collected by either Sheffield (N=1008) or Manchester (N=943) as these data sources generally genotyped the majority of GoRA subjects. Any subjects who had <50% of the SNPs present were removed from the dataset leaving N=912 subjects for analysis. Any SNPs which had no present results for subjects were also removed resulting in 368 SNPs for analysis. This dataset would be suitable to develop modelling methods which can be used on larger datasets. The issue of missing data was re-examined in section 4.2.3.

The second dataset hereafter called the 'GWAS SNPs' dataset, consisted of SNPs collected by any source but only retained for subjects who were in the Milan GWAS source data (N=394). This is the main dataset for the research as it enables the selection of a few important SNPs from thousands of presumed unimportant ones.

SNPs were coded in terms of number of copies of the minor allele (i.e. 0, 1 or 2). This is consistent with approaches by other researchers analysing high-dimension data (Le Cao et al., 2011, Dimauro et al., 2011, Long et al., 2011, Wang et al., 2009, Le Floch et al., 2012).

2.2.4. GoRA study data summary

The patient data which is available is cross sectional with varying degrees of disease duration. The Larsen score is a validated measure of erosive joint damage. It is considered the most robust measure of severity which is not subject to fluctuations over time (monotonic). This will be the main measure of severity for modelling. In addition, multiple severity variables will be modelled together as severity of RA can be measured in many ways. PCA analysis (section 2.2.1.2) of the severity measures identified the following groups to model based on their correlation.
1) 8 domain SF-36 variables
2) 4 SJC/TJC variables and DAS28
3) DAS28, PVAS, RASEV, MHAQ, ESR, CRP, Any Erosions and the Larsen score (including the separate hand and foot counts)

As a result of the Milan GWAS not being available for all subjects, two sets of data will be investigated. Firstly, the 'all subjects' dataset will be used to develop the methodology (Chapter 4). This dataset consists of all subjects (N=912) and 368 SNPs which were measured on the majority of GoRA subjects. Secondly, the methods will be applied to a much larger dataset (Chapter 6) consisting only of the subjects who were included in the Milan GWAS (N=394). This 'GWAS SNPs' dataset after some data cleaning contains 325,482 SNPs. Demographic and environmental data will also be included in both datasets and are described in sections 4.2.2 and 6.2.2.

2.3. Literature review of genetic variants contributing to the severity of RA

2.3.1. Introduction and methods

There has been substantial research using large cohorts investigating the genetic risks contributing to RA susceptibility (Eyre et al., 2012, Stahl et al., 2010, Okada et al., 2012). However, to date, studies investigating the genetics of RA severity have consisted of much smaller studies. This is probably due to cohorts with the more complex measures of severity not being available.

Marinou et al. (2010) performed a systematic literature review of articles published before November 2008 which explored the genetic influence on RA severity. The search was performed using PubMed using the following terms: 'Genetics' and 'Rheumatoid arthritis' and 'Radiographic damage' or 'Radiological damage' or 'Severity'. The findings by Marinou et al. (2010) were combined with a Medline systematic literature review completed on the 7th December 2010 (full results shown in Appendix B) followed by an updated review on the 14th June 2013.

The search criteria were widened to include other terms which could suggest radiographic or radiologic damage. The following search terms were used: 'rheumatoid arthritis' had to be in the title and the paper had to include the term 'severity'. In addition, there had to be at least one of the following terms 'Larsen score', 'Sharp score', 'bone erosion', 'bone resorption', 'erosion', 'radiographic*', 'radiologic*' or 'gene*'. The asterisk was used as a wild card to allow for all variations of ending the word. The initial search resulted in 265 papers being selected with a further 242 papers examined in the updated search. The literature review excluded papers investigating the effects of protein expression, consanguinity or the effect of medication on RA severity. It was decided that the exploration of these variables would not be possible with the

available data for this research. In addition to the systematic review, any papers which were found during general reading with evidence of genetic influence on RA severity were also included.

Of the identified papers in the searches, 75 were investigating SNPs or genes to predict severity of RA. The selected papers were assessed for quality using an adaption of the Critical Appraisal Skills Programme (CASP, 2010). Overall, it was felt that the quality of studies was good and they had mostly adjusted for confounding and attempted to minimise bias. However, many studies had very small sample sizes which could limit the generalisability of the results.

2.3.2. Results of literature review

The data from Marinou et al. (2010) was compiled along with the 75 papers selected in the later searches. This increased the genetic variants under investigation from 24 reported by Marinou et al. (2010) to 53 genetic variants. Any variants reported only as negative findings were not included. Appendix B presents alphabetically, all genetic variants found to contribute to RA severity in the systematic literature review (with updates made following the updated search). The sample size of the study, which severity measures were investigated, what analysis was performed and what the results were, are presented for each variant.

Due to the large quantity under consideration with inconclusive evidence, only the variants with the strongest evidence are reported below. As studies varied in size, quality and what was investigated, it was quite subjective to determine which variants could be considered likely to be truly associated with RA severity. It was decided that a variant had sufficient evidence, if there were two or more good sized cohorts (>150 subjects) reporting associations with RA severity and only either small sized cohorts or cohorts looking at different endpoints, subgroups or populations which contradicted the evidence. These variants are presented below in order of the variants with the largest number of cohorts finding evidence and the smallest number of cohorts published which found no evidence. Not all papers reported the size of effect, hence in some cases, only p-values and the study sample size are presented.

2.3.2.1. HLA DRB1

As described in section 2.2.3.1, there is substantial evidence in the literature of HLA-DRB1 (Shared epitope) association with disease activity and severity of RA (Marinou et al., 2010, Farouk et al., 2009, Mewar et al., 2008, Min et al., 2010, Mackie et al., 2012a, Meyer et al., 2011, Gyetvai et al., 2010, Tezenas du Montcel et al., 2005, Michou et al., 2006). Many of these authors report that a K-R-A-A amino acid sequence at positions 71-74 (S2 category) corresponds to the highest risk of RA susceptibility and severity. However, results for the other shared epitope categories can vary depending on the endpoint used to quantify severity, the coding used for the shared epitope and whether ACPA status is taken into consideration (section 2.2.3.1). One recent Japanese study concluded an 11.6% (95% CI: 4.1%-18.5%, p=0.0021, N=830) increase in joint damage, as measured by the Sharp score five years after disease diagnosis, for each copy of the RAA shared epitope motif that a patient has (Suzuki et al., 2013).

2.3.2.2. Peptidyl arginine deiminase, type IV (PADI4)

Peptidyl arginine deiminase, type IV (PADI4) showed association with disease severity in four separate studies. Suzuki et al. (2013) reported a 7.3% (95% CI: 0.14%-15%, p=0.037, N=830) increase in joint damage, as measured by the Sharp score five years after disease diagnosis, for each copy of the rare T allele. Hoppe et al. (2009) found a higher Steinbrocker score with rs2240340 T allele (p<0.004) and TT genotype (p=0.008) (N=373). Harris et al. (2008) reported serum levels for the Isoform 4 of the human peptidylarginine deiminase (hPAD4) to be associated with the RA susceptibility haplotype of PADI4. They reported subjects who were anti-hPAD4 positive (positive for autoantibodies) to have significantly worse Sharp score (P<0.001) (N=129). In support of this, Halvorsen et al. (2009) found anti-hPAD4 positive compared to negative patients, had a worse DAS28 (p=0.049), Sharp score (p=0.047) and change in Sharp score (p=0.023) (N=40). Just one study of 1384 Japanese patients was found in the review which did not replicate the above findings (Nishimoto et al., 2008). In conclusion there is strong evidence of an association between the PADI4 gene and RA severity but possibly not in the Japanese population.

2.3.2.3. Fc receptor-like protein 3 (FCRL-3)

Four recent studies have all concluded that the Fc receptor-like protein 3 (FCRL-3) -169 T>C polymorphism (rs7528684) is associated with erosive RA. Han et al. (2012a) reported higher Sharp scores associated with the CC genotype for patients in the ≥10 year disease duration subgroup (p=0.034, N=227). Maehlen et al. (2011) also found the CC genotype to be associated with 10 year radiographic progression (N=652). Chen et al. (2011) reported an increase in CC+CT genotypes for patients with destructive disease compared to non-destructive disease (Odds ratio [OR]=1.672, 95% CI: 1.149-2.432, p=0.007, N=670) which was replicated by Bajpai et al. (2012) (N=51).

2.3.2.4. Transforming Growth Factor Beta (TGFβ)-509, TGFβ+869 and TGFβ+915

Four studies have found RA severity associated with Transforming Growth Factor Beta (TGFβ) -509 (rs1800470), TGFβ+869 (rs1800469) or TGFβ+915 (rs1800471). Ceccarelli et al. (2011) (N=77) reported the TGFβ+869 TT genotype was associated with a lower MTP joint total erosion score (TT genotype mean (STD) = 6.3 (5.78), CC/CT genotype mean (STD) = 11.7 (7.8), p=0.011). Mattey et al. (2005) (N=208) detected a higher mean health assessment questionnaire (HAQ) score (p=0.04), higher Larsen score (not significant after adjustment for disease duration) and higher mortality (p=0.01) associated with the T allele at +869. Kim et al. (2004) found a significantly higher Sharp score associated with the T allele at the -509 position (p=0.048, N=143) and Oen et al. (2005) reported the homozygous TGFβ1 codon 25 G/G genotype (TGFβ+915, rs1800471) being protective against joint space narrowing after two years (OR=0.176, 95%CI: 0.037-0.837, p=0.029, N=181).

2.3.2.5. IL-6 -174 rs1800795 (possibly in RF or ACPA positive patients only)

There was mixed evidence of an association of interleukin (IL) 6 -174 (rs1800795) and disease severity. Marinou et al. (2007) (N=964) concluded a significant difference in modified Larsen score (median CC genotype=25, CG=27 and GC=33.5, p=0.005) however when analysed separately by ACPA and RF status, the association was only found in patients who were either RF positive (p=0.004) or ACPA positive (p=0.01). Although no other studies were found reporting an association with radiographic severity, Pawlik et al. (2005c) reported a more active disease with the GG genotype compared to GC and CC as measured by DAS, ESR, TJC and SJC (N=98) and Oen et al.

(2005) found a positive correlation between the GG genotype and pain (β= 0.899, 95% CI: 0.185-1.612, P = 0.014, N=181). Although Oen et al. (2005) also investigated radiographic damage they found no evidence of an association.

Ceccarelli et al. (2011) reported that the MTP joints total erosion score GC genotype mean (STD)= 7.4 (8.1) was significantly less severe than the CC genotype mean (STD)=10 (8.8), p=0.007, N=77). Although this suggests less severe disease with the G allele, the other joints in this study were not significant and comparisons of GG versus CC were not significant. Therefore, this could be a false positive.

Two studies were found investigating the serum levels of IL-6. Gottenberg et al. (2012) (N=578) found higher levels of IL-6 were associated radiographic progression at one year (OR =2.4, 95% CI: 1.1 to 5.2, p=0.005). Lamas and Rodriguez-Rodriguez, have two papers (Lamas et al., 2010, Rodriguez-Rodriguez et al., 2011) concluding DAS28 and plasma sIL-6R levels are positively correlated with ACPA positive patients (r=0.45, p=0.0336) and negatively correlated with ACPA negative patients (r=-0.45, p=0.0825). In addition, they report an interaction between IL-6R rs8192284 and presence of ACPA for the DAS28 score (p=0.008, N=281) which supports the evidence of a link between ACPA positive disease and IL-6 reported by Marinou et al. (2007).

2.3.2.6. Chromosome 5 open reading frame 30 (C5orf30)

The T allele of Chromosome 5 open reading frame 30 (C5orf30) (rs26232) has recently been shown to be associated with a reduction in joint damage scores as measured by the Larsen score or Sharp score (Teare et al., 2013). Using 885 of the GoRA cohort subjects (as described in section 2.2), Teare et al. reported a median joint damage for the CC genotype of 31, for the CT genotype of 27 and for the TT of 16 ($p=4x10^{-4}$). To provide replication, a meta-analysis with two other cohorts (n=581 and N=418) found a severity ratio of 0.90 (95% CI: 0.84-0.96, p=0.004) associated with presence of the T allele compared to the C allele.

2.3.2.7. CD40 rs4810485 ACPA positive patients only

The genotype TT versus GT/GG of rs4810485 CD40 was found to be statistically associated with a higher rate of joint destruction in ACPA positive patients (van der Linden et al., 2009) (p=0.003, N=563) and this finding was successfully replicated in the same paper using a replication set of 383 patients (p=0.021).

2.3.2.8. Chemokine receptor type 5 (CCR5)

Two studies were found showing evidence of polymorphisms of the Chemokine receptor type 5 (CCR5) gene being associated with RA severity. Han et al. (2012b) reported a significant increase in total Sharp score associated with both the -1118 CTAT (insertion/deletion) in CCR5 (rs10577983) (p=0.048) and 303 A>G (rs1799987) (p=0.048) (N=357). In addition, when analysing the erosion score alone, there was an increase in the statistical significance (corrected for multiple testing p values of p=0.028 and p=0.028 respectively). Zapico et al. (2000b) also reported a CCR5-Δ32 gene association with severity of RA (non-severe RA vs. severe RA, N=160, p=0.012). However Graudal (2004) and Pokorny et al. (2005) reported no evidence of association.

2.3.2.9. Protein tyrosine phosphatase, non-receptor type 22 (PTPN22)

Although the protein tyrosine phosphatase, non-receptor type 22 (PTPN22) is well established as a predictor of RA susceptibility, its potential relationship with RA severity is less clear. Two studies (Lie et al., 2007, Marinou et al., 2007) found an association of PTPN22 (rs2476601) and radiological damage (P=0.01, N=238 and p=0.04, N=964 respectively). However a further six studies did not find any association (Pierer et al., 2006, Steer et al., 2005, Graudal, 2004, Karlson et al., 2008, Morgan et al., 2010, Innala et al., 2008).

To explore PTPN22 further and its potential relationship with smoking, a meta-analysis of six cohorts totalling 2680 RA patients to investigate ACPA status and eight cohorts totalling 3172 RA patients was performed to investigate presence of erosive damage. Both smoking and the PTPN22 genotype were found to increase the risk of ACPA positive disease both individually and in combination (OR=2.22, 95%CI 1.69-2.91, p=8.3 x10-9). However there was no evidence of an increase or decrease in risk of erosions despite association between ACPA positive disease and erosive damage (Taylor et al., 2013).

2.3.3. Summary of the literature review of genetic variants contributing to RA severity

Due to the large quantity of genetic research performed in this area, only the Medline database was searched which could lead to papers being missed.

It is immediately apparent that replication of findings in this research area has proven to be difficult. Most variants tested in more than one study have had contradictory results of the association between RA severity and the genetic variant. There are many factors which could explain why the studies fail to replicate findings:

1) Not all studies are looking at the same variant within a gene and so depending on linkage disequilibrium (LD) could be testing different associations.
2) Studies are not using the same outcomes measures. It is possible, that different genetic markers could have a different influence on disease activity compared to erosions.
3) Not all analyses took into account potential confounding factors such as ACPA status, RF status, time since disease onset or prior treatment for RA.
4) The type of analyses varied greatly and some analyses may not be appropriate as assumptions of normality and homogeneity were not explored.
5) Sample sizes vary greatly across the studies and because some SNPs have rare alleles there may not be the power to detect the smaller effects particularly when continuous data has been reduced to binary data for analysis.
6) Not enough power to investigate interactions and very few interactions investigated.
7) Studies followed patients for different follow up lengths.
8) Different ethnic populations were studied.
9) Often the size of the effect and precision could not be determined from the paper.

Despite the above, the following genetic variants were found likely to have an effect on RA severity in approximate order of strength of evidence.

- HLA-DRB1
- PADI4
- FCRL-3
- TGFβ-509, TGFβ+869 and TGFβ+915
- IL-6 -174 rs1800795 (possibly in RF or ACPA positive patients only)
- C5orf30
- CD40 rs4810485 (ACPA positive patients only)
- CCR5
- PTPN22 (possibly not for erosive damage but on ACPA positive disease)

In addition, the following genetic variants appeared in more than one study as having an effect on RA severity, however, it was felt more research was required to confirm the findings. For more details of the evidence for and against see Appendix B.

- Caspase recruitment domain family, member 8 (CARD8) TUCAN rs2043211
- Cyclin-dependent kinase-6 (CDK6)
- Cyclooxygenase 2 (COX-2) -765 (for patients without the shared epitope)
- Glutathione S-transferase Mu 1 (GSTM1) or theta 1 (GSTT1)
- IL-1α, IL-1β+3954 rs1143634 and IL-1β-511 rs16944
- IL-1RN IV, +2018 and Variable Number of Tandem Repeats (VNTR)
- IL-4 and IL-4 receptor
- IL-10 rs18000872 (possibly in ACPA negative or RF negative patients only)
- Mannose-binding Lectin (MBL) regions of defective 0/0 genotype and -221 for ACPA positive patients only
- Matrix metalloproteinase (MMP)-3 5A/6A rs3025058 and MMP-1 1G -1607 rs1799750
- TNFα -308 rs1800629
- TNFAIP3/ OLIG3 rs6920220 and rs10499194 (possibly for ACPA positive disease only in long standing RA)
- TNF receptor associated factor 1 rs10818488

2.4. Review of environmental factors contributing to the severity of RA

2.4.1. Introduction and methods

In addition to data on the genetic variants, the GoRA study is rich in the collection of environmental data. Therefore, a literature review was performed to investigate which environmental factors have been found to be important in RA severity prediction.

Papers identified in section 2.3 were combined with two Medline searches performed on the 8th July 2011 (search terms of Rheumatoid Arthritis, Sever* and predict* in the title) and 13th July 2011 (with the search terms of Rheumatoid Arthritis and Sever* in the title). This produced 21 papers reporting exploration into environmental effects on RA severity. This limited search may not have identified all papers exploring environmental factors likely to be predictive of RA severity however it does help to identify GoRA study variables which it is hoped are found to be important in the multivariate predictive modelling.

2.4.2. Results of literature review

2.4.2.1. Alcohol consumption

Maxwell et al. (2010) reported a decrease in RA severity associated with an increase in the frequency of alcohol consumption which remained significant after adjusting for age, gender and smoking status. Although other papers which investigated alcohol and RA severity could not be found, a systematic review and meta-analysis investigating susceptibility and alcohol by Scott et al. (2013), found a summary OR of 0.52 (95% CI 0.36, 0.76) for the reduction in risk of RA for drinkers versus non-drinkers.

2.4.2.2. Age

Bukhari et al. (2002) reported age at onset of symptoms (grouped into decades) to be predictive of increased severity of radiographic erosions along with RF status. Lodder et al. (2004) found age to be associated with low bone mineral density at the hip and spine and that low bone mineral density at the hip was associated with high Larsen scores for hands and feet. Hence they conclude age is associated with a more severe Larsen score.

2.4.2.3. BMI

Of the seven studies found to investigate disease progression and BMI, five studies reported lower BMI to be associated with higher radiographic joint damage with the studies investigating 2007 subjects in total (Joerg et al., 2004, Lodder et al., 2004, Caplan et al., 2013, Baker et al., 2011, Velpula et al., 2011). Two studies found a higher BMI associated with disease activity, however, one of these studies (Ajeganova et al., 2013, N=1596), did not measure erosions and the other study (Ibn Yacoub et al., 2012b, N=250), only found a weak positive correlation between high BMI and erosions ($r=0.297$, $p<0.001$).

2.4.2.4. Female hormones

Jorgensen et al. (1996) reported an increase in severity for patients who had more children, breast fed longer and breast fed more children. They also reported a protective effect of the oral contraceptive pill after adjustment for age, number of children and breast feeding. However, it is hypothesised that this reduction is only associated with reducing more severe forms of RA (Vanzeben et al., 1990). Pikwer et al. (2012) found females experiencing early menopause were more likely to experience a milder form of RA, however, they reported no substantial difference in severity for females who took the oral contraceptive pill or breast fed.

2.4.2.5. Gender

Although there is no evidence of gender being related to a more severe form of RA (Gossec et al., 2005), it was noted by Ahlmen et al. (2010) that men possibly overestimate their ability to function hence rate themselves lower when completing scores such of HAQ and DAS. This is a confounding issue which may result in women scoring higher on disease severity scores.

2.4.2.6. Smoking

There is substantial evidence of an association between smoking and increased frequencies of RF (Mikuls et al., 2008, Westhoff et al., 2008, Papadopoulos et al., 2005, Masdottir et al., 2000, Saag et al., 1997). Although Westhoff et al. (2008) detected no RF positivity/smoking association with a worse DAS28 or radiological outcome, numerous other studies have reported a link between smoking and many measures of RA severity (Mattey et al., 2002, Nyahll-Wahlin et al., 2009, Papadopoulos et al., 2005, Masdottir et al., 2000, Saag et al., 1997, Soderlin et al., 2011, Ruiz-Esquide et al., 2011).

Taylor et al. (2013) also found that subjects who had ever smoked were more likely to have ACPA positive disease which was enhanced with the presence of the PTPN22 polymorphism. Mattey et al. (2002) reported the smoking association may also depend on the polymorphism at the GSTM1 locus. Whilst this could not be fully replicated by Lundstrom et al. (2011), they did find GSTM1 to be a risk factor for ACPA positive disease in non-smoking females over the age of 60 and a protective effect of GSTM1 in ACPA negative disease in men (N=2426).

Smoking is therefore generally regarded as the most important environmental factor to date. Although one study was found to report a decrease in erosive disease associated with smoking (Salliot et al., 2011), all other studies report an increase in severity associated with smoking.

2.4.2.7. Socioeconomic status & deprivation

Although no significant difference was observed between socioeconomic status and erosive damage, there was a difference between all other clinical measures of severity (i.e. HAQ, SJC, TJC, VAS and DAS) (Massardo et al., 2012, N=1093). Mackie et al. (2012b, N=6298) also reported a positive correlation between deprivation and RF positivity after adjusting for smoking however they did not observe the same relationship between deprivation and ACPA positivity.

2.4.3. Summary of the literature review of environmental factors contributing to RA severity

It is hypothesised that smoking leads to a higher RA severity. In addition, a younger age at the onset of symptoms and lower BMI may also lead to an increase in RA severity. There appears to be a protective effect of alcohol consumption associated with lower RA severity. This review gives an insight into potential environmental effects on severity which will be examined in a multivariate way during this research.

2.5. Summary

The GoRA study data available for use on this project has varied disease duration from one year to 65 years. As a result of this they have very varied disease severity; from a Larsen score of zero (no bone erosions or cartilage damage) to a maximum Larsen score of 160 (13.6% have a Larsen score of zero). 73% of subjects are females. The mean age is 61 (range 20-92) and the mean BMI is 26 (range 12.5-50).

All subjects have information on their smoking and alcohol habits recorded, as well as general demographics and laboratory tests including autoantibodies. 20% are smokers with an additional 39% being former smokers. 68% drink alcohol and 10% of these exceed more than five drinks a day almost every day. Although the mean age at time of diagnosis is approximately in line with population estimates (mean of 47 years old), there is concern that the cohort have an increased number of patients (particularly males) with RA incidence at a young age compared to current population estimates (see 2.2.2.2).

Two datasets are defined for analysis; 1) The 'all subjects' dataset containing 912 subjects which have 368 SNPs chosen due to previous evidence of possible effects on autoimmune diseases, and 2) The 'GWAS SNPs' dataset containing 394 subjects which have 325,482 SNPs measured from either a GWAS study, a study investigating the MHC region or SNPs chosen due to previous evidence of possible effects on autoimmune diseases. Both datasets contain demographic and environmental variables as described in sections 4.2.2 and 6.2.2.

This data resource is not without limitations. Data is recorded at a cross-section in time and hence no measures of changes in disease activity or severity over time are available. The varied duration of disease and very little information about the treatments administered over that period may be problematic in trying to predict the severity.

The Larsen score was selected to be the key measure for this analysis due to it being relatively stable over time. Unlike measures of disease activity such as laboratory measurements which can fluctuate, the Larsen score tends to remain the same (if progression is controlled) or get progressively worse over time. Modelling methods will analyse a single severity measure (Larsen score), followed by multiple severity measures, grouped according to their similarity as observed in a PCA analysis (section 2.2.1.2). Grouping the Larsen score into categories (mild, moderate, severe) and using a binary measure of erosions versus no erosions were explored. However, it was decided to focus on the continuous measure. This ensured no loss of information. In addition, grouping the Larsen score would be difficult due to no consensus of a clinical definition for mild, moderate and severe.

Various genetic and environmental factors were identified through literature reviews. Where genetic variants were investigated in multiple studies, there tended to be at least some disagreement regarding their influence on RA severity. Determining a list of variants with sufficient evidence to be thought of as highly likely to influence RA severity was quite subjective. However, the following list had two or more good sized studies without substantial evidence against the findings: HLA-DRB1, PADI4, FCRL-3, TGF-β, IL-6 -174, C5orf30, CD40, CCR5 and PTPN22. By far the most conclusive environmental factor to influence RA severity was smoking. Other factors thought to be detrimental to severity were a younger age at the onset of symptoms and a lower BMI. In contrast, alcohol was found to have a protective effect.

Therefore, primary focus of the research is to form a prediction model of Larsen score severity (and separately of multiple severity variables) by reducing a large quantity of SNPs and environmental factors to those most predictive. The predictors of RA severity identified in the literature reviews will be compared against the variables selected in the final multivariate models created in this research.

3. Review of multivariate methodology

3.1. Aims

The aim of this chapter is to investigate multivariate data analysis methods and determine the most appropriate methods for use in this research. The chosen methodology is described in detail with description of its application, strengths and weaknesses.

3.2. Introduction

As summarised in section 2.5 this project is challenged with investigating a large quantity of SNPs and environmental factors to form a prediction model of a single Y variable (Larsen score) or groups of correlated Y severity variables. The desired methodology must be capable of being applied to the 'GWAS SNPs' dataset (as described in section 2.2.3.2.3) which has many more variables (325,482 SNPs) than observations (N=394).

One difficulty of analysing SNPs in the same statistical model is that they are correlated together. The reason for this is that some combinations of alleles (haplotypes) occur more often than would be expected by chance alone. In a population, after many generations, the frequencies of the occurrence of haplotypes should be equal to the product of the population allele frequencies. When this is not the case, the SNP loci are said to be in LD. For SNPs in LD, there is a non-random association between alleles and this leads to correlation between genotypes. Therefore the method chosen to create a prediction model of RA severity needs to be able to allow for many potentially correlated variables in the same model.

Many genotyping techniques exist, from investigating one SNP at a time, to high density chip microarrays which can simultaneously measure tens of thousands of SNPs. Retaining SNPs in a final prediction model which are not truly predictive of RA severity unnecessarily complicates the model. A variable selection method is required to only retain important variables predictive of RA severity in the final model.

Therefore, the aim of this project is to create a predictive model which analyses correlated variables in the same model and applies a variable selection technique to reduce the number of variables retained in the final model.

3.3. Multiple linear regression

Suppose a multiple linear regression model with a single response variable y and m independent variables ($x_1, x_2 \dots to\ x_m$). This model can be represented as:

$$Y = X\beta + \epsilon$$

Assuming n subjects, Y is a column vector (n x 1), X is a matrix (n x m), β is a column vector (m x 1) and ϵ is the residual error vector (nx1) assumed to be multivariate normal with diagonal covariance matrix.

The common method to estimate the β coefficients in order to establish a solution to the model is to use the "least-squares method" by solving $\beta = (X'X)^{-1}X'y$.

This identifies the first challenge with multivariate data as the inverse of X'X may not exist.

If there are more independent variables than the number of subjects measured (i.e. m ≥ n), then all variation in y can be explained by the model and there is no residual error (ϵ). If m > n, there are infinite solutions for β and the estimates for β using the least squares method becomes unreliable. In these situations, a model containing many variables may be able to predict the current patient's RA severity exactly, however may not be transferable to an independent set of data. Therefore, to form a reliable model, the number of variables to keep in the final model needs to be reduced to those which are truly predictive of y and not simply explaining random noise in the data. Where too many variables are included in a model, it is often referred to as "over fitting" which is further described in 4.2.1.2.

Determining which variables should be kept in the model can be decided using the size of the standardised β coefficients (or p-values). However, a further complication of multivariate data arises when modelling correlated variables, as the least squares method solution for the estimated coefficients of the model leads to the parameter estimates becoming unstable (Cox, 2005 p.190, Abdi, 2010, Geladi and Kowalski, 1986).

Any methodology able to overcome the challenges described above also needs to be able to model more than a single dependant variable (y) at a time.

3.4. Penalised multiple linear regression

To resolve the issue of too many variables in the model, a variable selection method is required. Using "forward selection", variables are entered one at a time into the model and the lowest p-value is used to determine the first variable to be included in the model. The process is repeated including the 1st selected variable and adding in the remaining variables one at a time. The variable with the next lowest p-value is then selected to accompany the 1st variable in the model and the process is repeated until no further variables meet a pre-specified threshold. One problem with using this method on SNP data is that it would be extremely time consuming, as variables are only entered one at a time and many models have to be fitted to achieve the final model. A bigger issue is that as SNPs are correlated, the p-value becomes unreliable, hence the order variables are entered into the model, or slight changes in the analysis sets of patients may result in one correlated variable being selected above another. The result is the production of very different models being created with no indication of which variables are preferential.

An alternative is to use a regression penalisation technique such as Ridge regression (L2 penalty), Least Absolute Shrinkage and Selection Operator (LASSO) (L1 penalty) or Elastic nets (a combination of L1 and L2 penalties). These methods use a "penalised" least squares method to estimate β instead of the least square method described in 3.3. The penalisation terms which are minimised in each of the three methods are shown below (Li and Sillanpaa, 2012):

Ridge regression: $\sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j=1}^{m} x_{ij}\,\beta_j) + \quad \lambda \sum_{j=1}^{m} \beta_j^2$

LASSO: $\sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j=1}^{m} x_{ij}\,\beta_j) + \quad \lambda \sum_{j=1}^{m} |\beta_j|$

Elastic Net: $\sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j=1}^{m} x_{ij}\,\beta_j) + \quad \lambda\,[(1-\alpha)\frac{1}{2}\sum_{j=1}^{m} \beta_j^2 + \alpha \sum_{j=1}^{m} |\beta_j|]$

λ is referred to as the shrinkage factor and is pre-specified to determine the amount of shrinkage (reduction in the β coefficients) required for the model. The LASSO has the added advantage of shrinking some of the coefficients to exactly zero which has the effect of removing hopefully unnecessary variables from the model. Although Ridge regression shrinks the β coefficients it does not set them to zero, hence all variables are still retained in the final model even if they have small coefficients. Li and Sillanpaa. (2012) report that the LASSO has two disadvantages over Ridge regression. Firstly, when modelling highly correlated variables the LASSO tends to only select a single variable from the highly correlated group. In addition, the maximum number of independent variables which can be included in the model is restricted to n, the number of observations.

By combining the L1 and L2 penalties, Elastic net overcomes the limitation of only being able to select n independent variables (restriction in LASSO) and can still tend some β coefficients to exactly zero (restriction in Ridge regression). When α=0, the Elastic net equation resolves to ridge regression. When α=1, it resolves to the LASSO and for all other values ($0 < \alpha < 1$) it uses a combination of both L1 and L2 penalties.

Much research using these types of methods has been conducted recently, mostly using penalised logistic regression to predict susceptibility from genetic variants (Ayers and Cordell, 2010, Abraham et al., 2013, Li and Sillanpaa, 2012). One problem of using these methods is that they are still governed by standard linear regression assumptions (linearity, normally distributed independent errors and constant variance).

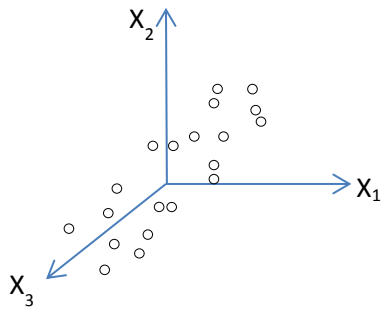
Whilst it was felt these methods could have been examined in this research if time permitted, there were some concerns over their application. Firstly, it was felt that the distribution of the Larsen score may make it difficult to find an appropriate model (although zero-inflated negative binomial [ZINB] models could be considered). In addition, the choice of λ and α for each model may heavily influence the variables retained for the final model and finding the optimum selection could be very time consuming on such a large set of data. The main concern however was how the models would handle the large quantity of correlated variables and whether the beta coefficients estimated would be reliable and reproducible.

One solution would be to pre-filter the SNPs, selecting one variable to represent a group of heavily correlated variables prior to modelling. Whilst this successfully reduces the degree of correlated variables entered into the model and makes the beta coefficients more stable, Abraham et al. (2013) recently warned against the use of pre-filtering as it could exclude important signals in the data early in the modelling process. It was therefore decided to try to find a method which does not exclude SNPs until their relationship with the RA severity variable had been considered.

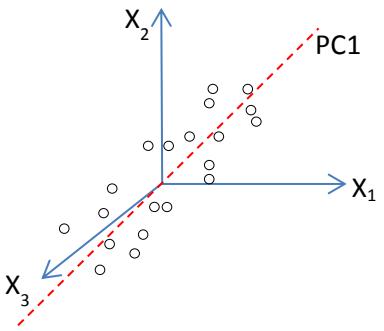
Methods involving dimension reduction overcome the correlated variable problem by projecting the original variables onto a plane. In general, a linear transformation is performed on the original correlated set of variables to produce an independent set of components which can then be used in the modelling. Probably the simplest form of this is PCA which is fully described below to aid later discussion on more advanced methods.

3.5. Principal components analysis (PCA)

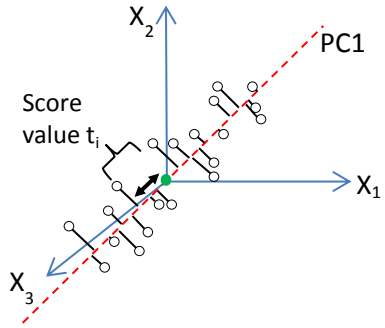
A set of correlated variables can be written into a matrix X. Using PCA this matrix is rewritten as a sum of further matrices. For example: $X = PC_1 + PC_2 + PC_3 + \ldots + PC_h$. Where X is a (nxm) matrix, and each PC_h is a (nxm) matrix. Each matrix PC_h can be further written as the outer product of two vectors, a score t_h (nx1) and a loading p'_h (1xm). Such that $X = t_1 p'_1 + t_2 p'_2 + t_3 p'_3 + \ldots + t_h p'_h$. Visually, Eriksson et al. (2006a, p. 46) and Geladi and Kowalski (1986) use diagrams similar to those shown in Figure 3.1 to demonstrate the relationship.



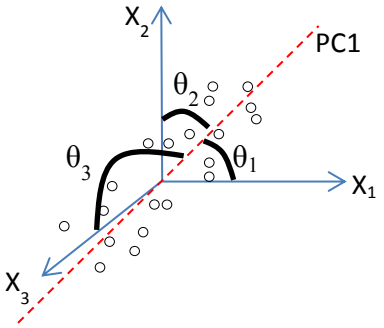
1. After mean centring and scaling to unit variance, the data for 20 subjects are plotted with each axis representing one of three X variables (X_1, X_2 and X_3).



2. The 1st component (PC1) is computed as the line corresponding to minimising the sum of the squared residuals (least square error). The line must go through the origin (mean of the data).

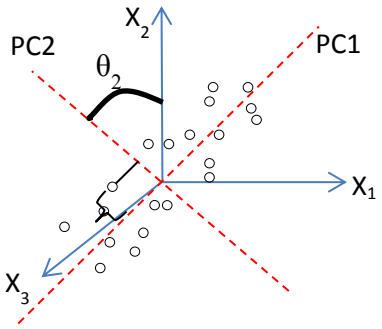


3. Each subject (i) is projected onto the PC1 line (forming a 90° angle) and the distance from the origin to each subject's projection is calculated. The distance is the score value t_i, a value for each subject.



4. Each variable (X_j) is rotated onto the PC1 line. The cosine of the angle θ_j equals the loading (p_j). Hence, a variable close to the PC1 line will have approximately cosine (0) or cosine (180) =1 (a high loading value). A variable almost orthogonal to the PC1 line will have approximately cosine (90) = 0 (a low loading value).

Figure 3.1 Principal component analysis



5. After the first PC1 line has been fitted and the scores and loadings calculated, the second component (orthogonal to the first can be fitted and the second set of scores and loadings calculated in the same way.

6. The common method for estimating principal components is to use Non-linear iterative partial least squares (NIPALS). This algorithm calculates t_1 and p'_1 from the X matrix. Then the residual error is calculated as $X - t_1 p'_1$. This residual is then used to calculate t_2 and p'_2 and so on until all components are calculated. The NIPALS algorithm can be found fully described in Geladi and Kowalski (1986). A summary is reproduced here where h is the index for the component being calculated.
Take a column vector X_j: $t_h = X_j$
Calculate the row vector loadings:
$p'_h = t'_h X / t'_h t_h$
Normalise p'_h to length 1: $p'_h = p'_h / || p'_h ||$
Calculate $t_h = X p_h / p'_h p_h$
Compare the t'_h used in step 2 with that obtained in step 4. If they are the same (within a criterion) then stop, else go back to step 2 using new t'_h.
Continue to calculate subsequent components but replacing X_j in step 1 with the residual error not yet explained by the calculated components.

Figure 3.1 Principal component analysis (continued)

3.6. Principal component regression

Following the use of principal component analysis to remove correlation in the variables, the resulting principal components which are by definition uncorrelated, could be used in a regression. This is called Principal component regression. In the discussion below, the X and Y data are assumed to be mean-centred and scaled prior to analysis. Firstly, let's consider the case of modelling a single response variable y. X is a matrix (nxm), y is a vector (nx1), n=number of observations, m=number of X variables.

The first component is calculated for the X variables using a PCA analysis as described in section 3.5 and demonstrated again in Figure 3.2. This provides the scores (t_1) and loadings (p_1) for the first component which are used to represent the X data: $X = t_1 p'_1$. Subsequent components can be calculated as described in Figure 3.1.

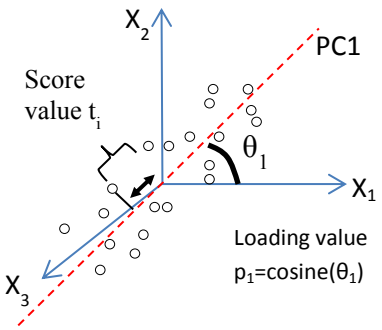


Figure 3.2 Principal component analysis – First component

In the single y case, the scores (t_1) are used with the scaled y data and b_h coefficients calculated using regression $\hat{y} = b_1 t_1$ where $b_1 = y' t_1 \,/\, t_1' t_1$ (Figure 3.3). By reducing the dimension of the data prior to the regression and using the t scores instead of the X data, the linear regression assumption of independent errors should now hold, assuming subjects are independent.

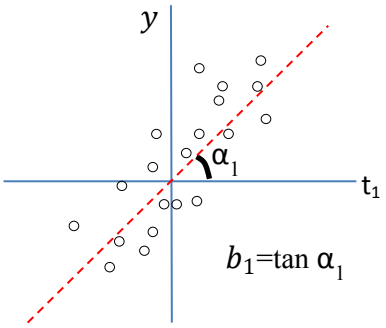


Figure 3.3 PCA regression of y= t scores

Whilst PCA regression removes the problem with unstable beta coefficients (as the principal components used in the model are uncorrelated), the components are derived using all variables in the model because the PCA is performed in isolation to the regression. Therefore there is no opportunity to reduce the number of variables needed to be measured in order to use the prediction model. A further concern is that all of the components are calculated without reference to the y variable and hence it may not perform as well for prediction as selecting variables based on their predictive importance.

3.7. Partial Least Squares (PLS) regression

Initial exploration into statistical methods each revealed problems with modelling multivariate data where the number of variables is substantially larger than the number of observations (m>n). Early exploration into using multiple linear regression methods revealed unstable β coefficients when modelling correlated variables together. When using 912 subjects' data, it was found that no more than 30 variables could be fitted in the same model before model convergence problems were observed. One solution is to test SNPs one at a time in the model, however, this has the potential of detecting false positives due to multiplicity and is very time consuming, when thousands of separate single SNP models are run.

Whilst PCA regression made positive steps towards removing the collinearity and dimensionality problem, the components to represent the X variables are defined unsupervised on the Y variable. Partial Least Squares (PLS) is a natural extension from these methods. It simultaneously applies dimension reduction in the general form of a PCA analysis on both the sets of X and Y variables and performs a regression using the calculated scores and loadings. Unlike PCA regression, the PLS algorithm performs dimension reduction and regression simultaneously ensuring the selected components are optimised for Y variable prediction. Further discussion is summarised in section 10.3 as to why PLS is a good choice compared to other methods. In addition, section 3.8 describes other methods which were investigated before deciding to use PLS as the main focus of this research.

3.7.1. Methodology

The PLS algorithm as described by Geladi and Kowalski (1986), annotated for having a single Y variable, is described below. Notice how unlike the PCA algorithm (Box 6 in Figure 3.1), the Y data is used in the calculation of the X components (scores and loadings) and the X data is used in the calculation of the Y components (scores and loadings). For each component:

	1) Let u = y	u *is a vector (nx1)*
In the X block	2) w' = u'X/u'u	*w' is a vector 1xm (as u' is 1xn and X is nxm)*
	3) w' = w'/||w'||	(normalisation of w')
	4) t = Xw/w'w	*t is a vector nx1 (as X is nxm and w is mx1)*
In the Y block	5) q' = t'Y/t't	*q' is a scalar (as t' is a vector 1xn and Y is vector nx1)*
	6) q' = q'/||q'||	(normalisation of q')
	7) u = Yq/q'q	*hence u is a vector nx1 (as y is nx1 and q is a scalar)*
Check convergence:	8) compare the t in step 4 with the one in the preceding iteration. If they are approximately equal then go to step 9, else go to step 2.	
	9) p' = t'X/t't	p' is a vector 1xm (as t' is a vector 1xn and X is nxm)
	10) p'_{new} = p'/||p'||	(normalisation of p')
	11) t'_{new} = t' ||p'||	(rescaling scores)
	12) w'_{new} = w' ||p'||	(rescaling weights)
Find b (a scalar as u' is 1xn and t is nx1)		
	13) b=u't/t't	

After the scores (t) and loadings (p) for the first component are finalised, (shown as t'_{new} and p'_{new} above), the information not yet able to be predicted by the model (residuals) are calculated and used to replace X and Y in the equations above when calculating the remaining components. For example, for h components the X data residuals are calculated as using the finalised scores and loadings for the relevant component:

$E_h = E_{h-1} - t_h p'_h$; $X=E_0$ where h is the number of components.

Similarly, the Y data residuals (for component h) are calculated using: $F_h = F_{h-1} - b_h t_h q'_h$; $Y=F_0$

Note that because $b_h t_h q'_h$ is used to calculate the Y residual as opposed to $u_h q_h$, the rank of Y is not decreased after each component and m-1 components can be calculated.

Once the X in steps 2, 4 and 9 and Y in steps 5 and 7 are replaced by their corresponding residual matrices E_h and F_h, the algorithm returns to step 1 and the loadings and score for the 2nd component are calculated.

In the case of multiple Y variables where Y is a matrix (nxc), steps 1 to 13 become:

For each component:	1) Let u = y_k	u *is a vector (nx1), the k^{th} y variable used to start*
In the X block	2) w' = u'X/u'u	*w' is a vector 1xm (as u' is 1xn and X is nxm)*
	3) w' = w'/||w'||	(normalisation of w')
	4) t = Xw/w'w	*t is a vector nx1 (as X is nxm and w is mx1)*
In the Y block	5) q' = t'Y/t't	*q' is a vector 1xc (as t' is a vector 1xn and Y is vector nxc)*
	6) q' = q'/||q'||	(normalisation of q')
	7) u = Yq/q'q	*hence u is a vector nx1 (as y is nxc and q is cx1)*
Check convergence:	8) compare the t in step 4 with the one in the preceding iteration. If they are approximately equal then go to step 9, else go to step 2.	
	9) p' = t'X/t't	p' is a vector 1xm (as t' is a vector 1xn and X is nxm)

10) $p'_{new} = p'/||p'||$ (normalisation of p')

11) $t'_{new} = t' \, ||p'||$ (rescaling scores)

12) $w'_{new} = w' \, ||p'||$ (rescaling weights)

Find b (a scalar as u' is 1xn and t is nx1)

13) b=u't/t't

Boulesteix (2004) and Boulesteix and Strimmer (2007) argue that PLS is the only known well-established dimension reduction method which works when the number of subjects is less than the number of variables, makes no distribution assumptions and chooses the reduced component dataset using both response and explanatory variables. This 'supervision' of X on Y, explains why PLS often performs better than principal component analysis in prediction problems. PLS can cope with many, collinear, noisy and moderately incomplete data for both the X and Y variables (Eriksson et al., 2004, Trygg and Wold, 2002) making it an attractive option for this research.

3.7.2. Extensions using penalised PLS (Sparse PLS [SPLS])

When a particular linear model is chosen, all variables in the model are assumed to have some predictive ability. However, in the 'GWAS SNPs' dataset which has 325,482 SNPs, it is very likely that there are many SNPs unrelated with the severity of RA. A final model should only retain SNPs contributing to the underlying RA severity and exclude SNPs that describe random error in the sample. As SNPs are being selected from over 325,000, it is possible to add variables to the model until every patient's severity score is perfectly explained (predicted correctly). However, this would be an over fitted model and it would have poor prediction on an independent cohort. See section 4.2.1.2 for further discussion on over fitting.

Therefore, a dimension reduction technique (which removes collinearity problems) and a variable selection technique (which removes over fitting of the model) is required.

PLS regression is a dimension reduction technique which solves the multiple linear regression problem of calculating the inverse solution to $(X'X)^{-1}$ in the presence of correlated variables. However, the default PLS algorithm does not have a variable selection technique and all variables are included in the final model (with all variables contributing to the loadings). One way to overcome this is to apply a penalised regression type strategy within the PLS algorithm. Sparse PLS (SPLS) provides a dimension reduction and variable selection method using a LASSO type penalisation (as described in section 3.4) alongside the standard PLS algorithm. For comprehensive descriptions of SPLS modelling see Chun and Keles (2010), Le Cao et al. (2008) and Feng at al. (2012). A brief description of the Le Cao et al. (2008) methodology is provided below and is easily implemented using the R package mixOmics, version 3.0 (González et al., 2011, Lê Cao et al., 2009). Other SPLS available software is discussed in section 3.7.5.

The PLS algorithm as described in section 3.7.1 is used to calculate the scores and loadings for the first component. Prior to finalisation of the first component, the loading vector is sorted by descending magnitude. As part of the model specification, the number of variables required in the final model for the X's (and Y's in the case of multiple Y variables) is pre-specified. Any variable, whose corresponding loading is not in the top required number, have their loadings set to be 0. Hence, these variables do not contribute to the calculation of the first component. The PLS

algorithm continues recalculating the final scores and loadings based on the selected variables. The residual error is calculated by subtracting the predictive ability of the 1st component, (using the reduced set of variables), from the original X and Y variables and this residual error is used as the basis to model for the 2nd component. The same process is followed for subsequent components up to the number of components required.

As SPLS appears to provide a modelling solution to the research question, it was decided to investigate it further in section 4. For the final model, the number of components retained requires careful consideration. Components explaining only a small amount of variation may only be describing random noise in the data and hence should be excluded from the final model (Geladi and Kowalski, 1986). This is further discussed in section 4.2.1. In addition, the pre-specification of the number of variables to keep in the final model also requires careful consideration. Further discussion is provided in 4.2.1.4.

3.7.3. Regression coefficients in PLS regression

The regression coefficients β in PLS regression are transformed back to the original scale (not scaled or with mean centring for interpretation). Whilst these β coefficients can be directly interpreted as the direction of effect that a variable has on the model, extreme care needs to be taken. If two 100% correlated variables are in the same model then the value/size of their β coefficients will be halved. Each variable will therefore appear to have a small effect whereas interest is in the total effect of both variables. An example of this can be found in SNPs rs10506802 and rs7968671 in Table 6.6. Unlike ordinary least squares regression whose βestimates become unstable in the presence of highly correlated variables, both variables are retained in the model, however, because they are 100% correlated, the estimates of their β coefficients will be halved. As these regression coefficients are not being used for variable selection (loadings are used instead), the fact that they are smaller in the presence of correlated variables is not a problem.

3.7.4. Advantages and disadvantages of PLS regression

Fornell and Bookstein (1982) and (Vinzi et al., 2010 p.659) explain that unlike covariance-based maximum likelihood estimation which assumes a joint multivariate distribution and independent observations, PLS does not require these distributional, population or scale of measurement assumptions. A possible limitation though is highlighted by Eriksson et al. (2006a) who warns that the modelling works best when the data is "fairly symmetrically distributed and have a fairly constant error variance." This could result in poor prediction models for the Larsen score because of the inflated number of patients with zero erosions and non-normal distribution of the data.

One limitation identified by Chun and Keles (2010), is that PLS cannot overcome the problem of modelling with a dataset which has a too small sample size. A reasonable number of observations are required in order to estimate the relationships between the sample covariance consistently. The size required is dependent on the data structures, however, if the sample is not large enough to represent the population, then PLS cannot overcome this and it may result in poor prediction.

Given there are only 394 patients in the 'GWAS SNPs' dataset, it may not be sufficient to detect potentially small but important SNP effects on disease phenotypes. Although two analyses containing just 320 RA subjects used PLS to investigate disease activity relationships using grip force, walking speed and pain intensity using the Swedish TIRA project cohort, they did not include

genetic variants in this research (Thyberg et al., 2005, Bjork et al., 2008). They also did not attempt to create and test a prediction model and only used PLS to identify variables which contribute to RA severity using a variance importance plot to indicate significant predictors.

Investigation will be required into the effect of missing data by performing a sensitivity analysis using multiple models (section 4.2.3). Trygg and Wold (2002) suggest PLS can cope with moderate amounts of missing data (up to around 20%) based on the work by Rannar et al. (1995).

As traditional modelling methods are not being used, parametric approaches such as the use of confidence intervals and p-values to assess the importance of variables in the model can not be used. Measures of model fit (such as the Akaike information criterion or the Bayesian information criterion) which use maximum likelihood methodology also cannot be used. Exploration into assessment of the model fit and model validation techniques will be explored in this research.

A limitation of SPLS is that subjects with missing data would be excluded from the analysis during the various matrix and vector product calculations. Genotyping often leads to some samples with missing data, therefore all of the GoRA patients have at least one data point missing. This would lead to a substantial number of subjects or SNPs being excluded from investigation. During the development of PCA and subsequently PLS, Wold (1975) developed the NIPALS approach which can be used to impute data. This approach is fully described in section 4.2.2. Lê Cao et al (2009) recommend using NIPALS separately on the Y and X data before performing a SPLS analysis which results in no imputations being required during the modelling. Two forms of imputation 'quick' and NIPALS are investigated using the 'all subjects' dataset in chapter 4 and four methods are investigated using the 'GWAS SNPs' dataset in chapter 5.

Another concern is to ensure variables collected in different units are scaled. For example, SNPs are recorded as 0, 1 or 2, however, a variable such as smoking pack years can range from 0 to over 100. No preference (weight) should be given to one variable above another. Therefore, it is recommended to use auto-scaling (Geladi and Kowalski, 1986, Trygg and Wold, 2002). Auto-scaling is achieved for each variable by subtracting the mean from each result and dividing it by the STD. This results in each variable having a mean of zero and a STD of 1, hence all variables have equal importance irrespective of the units they are collected in. This can be performed prior to fitting SPLS or performed as part of the mixOmics function.

The use of PLS to investigate gene-gene and gene-environment interactions has previously been explored (Wang et al., 2009). PLS uses the correlations within the X and Y matrices to calculate the loading vectors. The correlations between X and Y are used, ensuring the final model maximises both the amount of variation explained within the X and Y matrices and the correlation between the X and Y matrices. Multiple components can be calculated; hence the first component could represent the main effect of a variable whilst later components could use interactions with other variables to explain further variation in the Y matrix. Therefore, Wang et al. (2009) claim PLS has power to detect X/Y association even in the presence of gene-gene and gene-environment interactions. However no further research was found supporting this claim.

3.7.5. Software choices for SPLS regression

Many software packages are capable of performing PLS which by definition includes the dimension reduction step. These include; MATLAB® version 7.13 R2011b (The MathWorks Inc., Natick, MA, USA), SIMCA® Umetrics AB Version 13.0 (Eriksson et al., 2006a, Eriksson et al., 2006b), SAS® Version 9.2 of the SAS System for Windows Copyright © 2002-2008 (SAS Institute Inc. Cary NC USA), R Foundation for Statistical Computing v2.13.1 (Vienna Austria), Tanagra v1.4.44 released May 14 2012 (Rakotomalala, 2005) and CoreExpress v1.0 (Magidson, 2011). However, only R software packages were found to apply the variable selection in the form of a sparse penalty. Other packages did have additional benefits to R and therefore have been used to answer specific questions in this research. A full review of software used is provided in section 10.6.

Two SPLS regression methods were identified (both using a type of LASSO-L1 penalty), as packages in the R Foundation for Statistical Computing, Vienna, Austria (version 2.13.1). The SPLS R package version 2.1-0 (Chun and Keles, 2010) reduces the variables to retain in the final model by optimising two tuning parameters (Kappa and Eta) to obtain the best model fit under cross validation (CV) (section 4.2.1). Kappa is the number of components to extract (section 4.2.1.3) and eta is a sparsity parameter (for the L1 penalty) between 0 and 1. There were concerns that optimising the parameters could be very time consuming when fitting models to such a large set of data. In addition, when the SPLS method is applied to the 'GWAS SNPs' dataset, R reaches its maximum memory limit which results in the data having to be modelled in 40 separate lower level blocks (as described in section 6.2). In this scenario, the tuning parameters recommend only keeping disease duration to carry forward to the higher level model. It was of concern that using the tuning parameters on the lower level models results in optimising the fit of a poorly specified model. This may lead to insufficient numbers of SNPs being carried forward to the higher level model which would miss variables with small effects but potentially important predictors of severity.

To prevent the loss of too many variables at the lower level models, it was decided the mixOmics package would be most appropriate for this research, version 3.0 (González et al., 2011, Lê Cao et al., 2009). This package allows the specification of the number of variables you want to keep. Hence, you can force the model to keep more variables than necessary from the lower level model and investigate optimising the model fit on the higher level model.

To ensure the R mixOmics package was correctly fitting PLS models, identical models were fitted in R and SAS/STAT software, Version 9.2 of the SAS System for Windows Copyright © 2002-2008 SAS Institute Inc. Cary, NC, USA. Both packages resulted in the same predicted model.

3.8. Other multivariate methods explored and modelling selection

Numerous other multivariate methodologies were explored and are briefly summarised in this section. However, none were believed to be as well suited to this research question as using PLS regression.

- Canonical correlation analysis (CCA) measures the relationship between two sets of variables (X and Y) by estimating linear combinations of each set which have maximum correlation with each other. A prediction model could be obtained by constraining each linear combination to be orthogonal to the previous combinations found, often resulting in numerous components (dimension reduction). Unless used with a variable reduction technique such as regularisation/ penalisation techniques (as described in section 3.4) the resulting model would consist of all variables in the original X and Y matrices. Even when used with a regularisation or penalisation technique, canonical correlation can require large computational power when the number of variables in X and/or Y variable are very large (Le Cao et al., 2009). A recent study (Le Floch et al., 2012) reported that CCA can be problematic when the sample size is substantially smaller than the number of variables under investigation and in these cases univariate filtering may have to be used. As described in section 3.4, it was hoped to avoid univariate filtering so that if a final suitably predictive model was found, it would have a higher chance of having the causal SNP retained in the model (if the causal SNP had been originally measured).
- Cluster analysis uses the variables to group observations into 'clusters' of the most similarity. Whilst this approach could be used to cluster patients into groups of similar severity, it would not help to create a severity prediction model.
- Correlated component regression (CCR) (Magidson, 2011) applies a regularisation algorithm to model multiple correlated variables (X) to predict an outcome variable (Y). The dimension reduction is derived selecting each component to maximise its ability to predict Y. Each derived component is a linear combination of the original X's variables, however, it can be used with a variable reduction 'step down' algorithm so that a reduced number of variables are retained for the prediction model. CCR is very similar to PLS, however, only implemented in the COREXpress software (Magidson, 2011) and hasn't been widely used in the literature. It was therefore decided not to focus on CCR initially, but to explore it at a later date if there was sufficient time.
- Discriminant analysis classifies an individual to one or more pre-defined groups. This research aims to use continuous measures of severity as agreed clinical definitions of mild, moderate and severe RA are not available and lead to arbitrary cut offs and loss of information. Therefore discriminant analysis is not applicable.
- Factor analysis is designed to find hidden variables (latent variables) which cannot be measured directly but are thought to exist. Similar to PCA (described in section 2.2.1.2), groupings are defined without variable reduction and without consideration of Y prediction. It therefore would not be the most appropriate technique to reduce the large quantity of X variables to form a prediction model of severity.
- Neural networks can use patterns and correlations in the data to build a model capable of prediction. Many types of neural networks can be applied and the optimum prediction can be based on either supervised or unsupervised learning. Depending on the model input, neural networks can incorporate non-linear and asymmetric relationships. However, neural networks may be unsuitable for modelling data with collinear variables and large quantities of variables because they rely on the inversion of the variance covariance matrix (Vinzi et al., 2010). Whilst it

is possible to perform a PCA or univariate pre-filtering before modelling a reduced set of components or independent variables using neural networks (Eriksson et al., 2006a), this may not be optimal for the GoRA data. The initial PCA is performed unsupervised on the Y response. Therefore, the variable dimension reduction is performed ignoring information regarding the relationship between the X and Y variables. In addition, unless neural networks are used with a penalisation approach, there is no variable selection technique applied. Therefore, if using PCA first, the components would require all variables measured for the final model to be fitted. This could be avoided using SNP pre-filtering, however, there is evidence that pre-filtering reduces the ability to recover causal SNPs (Abraham et al., 2013).

- Random forest is the method of fitting many decision trees to classify observations to categories. The method selects a random set of X variables and the classification tree is produced based on a randomly chosen training set of observations. This model is then validated on the left out observations. This approach is repeated numerous times and the most common classification is assigned. By categorising a continuous variable into many categories it can also be used for prediction of a continuous trait. At the start of the research project, there was no published evidence of random forests being used on GWAS data to predict continuous traits and no evidence that the method could analyse multiple Y variables. Hence, the method was not investigated in this project due to PLS appearing to fit the data better. However, of recent interest has been the use of random forests to detect SNP interactions from GWAS data. Winham et al. (2012) found interactions to be difficult to detect in the absence of large marginal effects. Boulesteix et al. (2012) also recently reported that caution was required if using the Gini variance importance measure as it favours SNPs with large minor allele frequencies. Hence, further exploration of random forests may be of interest.
- Structural equation modelling (SEM) is a general term to reflect a multivariate data analysis extension to generalised linear modelling. It is often used to test underlying latent variables which cannot be directly measured. A priori model is pre-specified which estimates how multiple variables are related using sets of linear equations. Using variances and covariances of the variables, this model is tested to determine if the parameter assumptions fit the data. The model may then be adjusted based on the results and the process repeated with new priori or concluded to be a good fit. Whilst this method could be applied to the GoRA data, it is believed that given the huge quantities of data this research is investigating, SEM may have computational difficulties defining the relationships between such a large set of data as identified by Wold (1985). Although with modern computers this may no longer be such a problem.

Although there are many options which could be explored to see if they could form a predictive model of RA severity, it was decided to focus the research on the use of SPLS.

3.9. Summary

In order to determine the most important genetic and environmental factors predictive of RA severity, a multivariate technique is required which has no distributional assumptions, can model multiple collinear variables together (through dimension reduction) and is able to select the most important variables to form a final model (through variable selection). Although many multivariate approaches could have been explored, it was felt that SPLS was the most appropriate. This was due to it being reported to cope well with very large datasets and because it is a commonly used statistical approach in many other areas. Other possibilities such as penalised multiple linear regression, PCA, Principal Component Regression, CCR and CCA, were either not capable of both dimension reduction and variable selection or they didn't apply the methods simultaneously which could lose important information about the relationship between the Y and X variables.

Although two SPLS macros are available in R, mixOmics was selected because it allows the number of variables to be retained for the final model to be chosen, rather than running optimisation strategies using tuning parameters which was anticipated to be time consuming.

A review of published literature was unable to find any evidence of PLS being used to model environmental and genetic variants to predict RA severity. However, 16 studies in non RA areas (all since 2005) were found analysing SNP data using PLS regression or discrimination. All studies (except one) were using less than 45,000 SNPs or pre-filtered SNPs to fewer than 45,000 SNPs prior to analysis. Le Floch et al. (2012) analysed 600,000 SNPs with just 94 observations however the results of this paper were not available until the majority of this research was completed. The PLS (and SPLS) model fitting strategies used in the literature are discussed further in section 4.2.

In order to test the feasibility of SPLS modelling on this data, model fitting strategies will first be applied to the 'all subjects' dataset (N=912) investigating prediction of a single Y variable from 368 SNPs (Chapter 4). Secondly, the process will be extended to the 'GWAS SNPs' dataset using data on 394 subjects and investigating 325,482 SNPs (Chapter 6). Finally the methods will be extended to the modelling of multiple severity variables (Chapter 9).

4. SPLS analysis of Larsen score –'All subjects' dataset

4.1. Aims

The aim of this chapter is to use SPLS on the smaller set of data (912 subjects, 368 SNPs and 19 environmental variables) in order to develop an optimal strategy to find the best prediction model for the Larsen score. The strategy to take forward has to be scalable so it can be applied to the 'GWAS SNPs' dataset. Therefore, decisions in this chapter will be made with consideration for the future GWAS modelling requirements. The investigation includes:

- Model creation strategies such as CV techniques to select the number of variables and components to avoid over fitting
- Imputation methods
- Issues with missing data
- Transformations of the Larsen score
- Predictive ability of the model
- Interpreting the model using variance partitioning

4.2. Methods used for the model creation

4.2.1. Model creation strategy

Multivariate methods often involve the use of an iterative process to determine which model fits the data the best (Eriksson et al., 2006a, p. 378). Boulesteix and Sauerbrei (2011) discuss numerous validation techniques in the context of high throughput molecular data. They highlight that to test the predictive ability of the model (and hence find the best model) the ideal scenario is to have an independent set of data available which is not used at all in the model creation. Unfortunately in practice, this is often not available and instead the existing sample can be randomly split prior to any analysis into a test and training sample. The size of the test and training sample varies, however, it is generally regarded that the training sample should be the larger (perhaps 80% or 90% of the data) so that the majority of data is used to form an accurate model and a smaller amount to test the model. If there are insufficient observations (subjects) to reserve sufficient data for testing, then they suggest using a CV technique.

Using the 'all subjects' dataset (N=912), 20% of the patients (N=182) could be reserved for a test set and the remaining (N=730) could be used as the training set to form the model. Whilst this seems sensible, an approach is required which would be applicable to the 'GWAS SNPs' dataset. As the 'GWAS SNPs' dataset has only 394 subjects, this would suggest 315 for the training sample and just 79 for the test sample. If a SNP had a minor allele frequency of 5%, this means on average, only four subjects would have this allele in the test sample. It is therefore likely that some SNPs found to be important in the training data would have no rare alleles in the test data. To retain highest chance of detecting important rare SNPs it was decided that there was insufficient data to keep a test set in reserve and CV would be used instead. This decision is reinvestigated in section 8.3 and 8.4.

4.2.1.1. Cross Validation

Many authors (Long et al., 2011, Le Cao et al., 2011, Daetwyler et al., 2013) use CV with SPLS for making decisions about the format of the model (choosing the number of components and number of variables which makes the model fit the best). CV can also be used to provide an estimate of the final model fit. Forming a model on a set of subjects and then using the model to predict the same subjects, will overestimate the predictive performance which would be observed on an independent set. CV can take many forms as described below, however, in all cases it uses some subjects to form the model and then uses this model to predict the subjects omited. Although it can provide better estimates of the prediction performance which may be observed on an independent set of subjects, the performance can still be over-estimated due to both creating and testing the model on the same set of subjects (optimisation bias) (Varma and Simon, 2006).

The most common CV methods in this field documented by SAS® (2008), Eriksson et al. (2006a, p. 374) and the package mixOmics (González et al., 2011, Lê Cao et al., 2009) version 3.0 are the leave one out method and the random block method. In the leave one out method, each subject is left out of the training sample one at a time and the model is then used to predict the left out patient. In the random block method, the data is split into M folds, the model is formed on M-1 of the folds forming the training group and tested on the Mth fold which was left out of the training sample. This is repeated until each of the folds have been the test group and all patients have been assigned a prediction.

The 'leave one out' method is better used when the sample is very small as at these times, (<20 observations), there is not enough data to use the random block method. At other times it is time consuming, as the number of models to be run, equals the number of subjects. R^2 is defined as the proportion of total variation in the Y response variable(s) explained by the model fitted on the entire set of data. When using the 'leave one out' method, each run contains a similar set of patients (the permutation of the data is insufficient) therefore Q^2 (R^2 under cross validation or R^2-CV) tends to approach R^2 and therefore may not be a reliable estimate of model performance on an independent sample. SAS® discusses other non-random methods of assigning blocks, sequentially (observations 1-50, 51-100 etc.) and split sample (observations 1, 21, 41 form a block and then observations 2, 22, 42 form a block etc.), however no evidence was found as to why this would perform better than the blocks being randomly assigned.

A further method which could be used in this context is bootstrapping. Although not initially investigated as it was not mentioned in the literature as being used with PLS modelling, bootstrapping is described and investigated on the 'GWAS SNPs' dataset shown in section 6.4.3.

Therefore, it was decided to use all of the data in model creation and to use internal random block CV methods to both select the best model and then to estimate how well the final model fits under CV.

4.2.1.2. Over fitted models

As described in section 3.7.5, care is necessary when modelling such a large dataset to avoid an over fitted model. This is a model which includes too many variables so that both the predictive signal and the random noise in the sample is explained by the model. An over fitted model will predict very well on the set of data which was used to create the model, but very poor on an independent set, in this case fewer variables in the model would have predicted the independent set better.

Two decisions which have to be made in SPLS modelling are the number of components and the number of variables to retain for the final model. If all possible components are included in a SPLS model, then the model will fit the data perfectly. However, the model would not be generalisable as the smaller components are explaining random noise in the sample and the model is over fitted. Similarly with 325,482 SNPs and environmental variables, a combination could be found which fits the data perfectly, however, the model would not be transferable to an independent set. Determining the optimum numbers of components and variables are described in sections 4.2.1.3 and 4.2.1.4.

4.2.1.3. Choice of the number of components

PLS models by definition extract the first component (a linear combination of the X variables) to represent as much of the Y variation as possible. Each subsequent component is formulated to explain as much as the left over residual variation as possible (section 3.7.1). Therefore the higher number of components that are fitted, the less of the original Y variable(s) variation it will explain.

As more and more components are fitted, they could simply be describing random error and not predictive residual left over in the Y's. Therefore these components which describe very little of the Y variation should be left out of the final model. For over 30 years, CV has been suggested as a technique to estimate the best number of components to use in a prediction model (Eastment and Krzanowski, 1982, Geladi and Kowalski, 1986). This is still the recommended approach used currently (Daetwyler et al., 2013, Long et al., 2011, Le Cao et al., 2011).

Using CV, the additional components added to the model are investigated to determine if they are of significant importance to include them or possibly just explaining noise (Sjostrom et al., 1983). The significance is estimated, for example in 4-fold CV, by deleting one quarter of the observations and then predicting the Y's for these missing observations. The sum of square (SS) differences between the observed and predicted Y values is calculated for each of the four quarters of data. This can be described as the prediction residual sum of squares (PRESS). The component associated with the minimum PRESS corresponds to the model explaining the most variation. Plotting PRESS by the number of components used, the minimum level can be observed and the component number this occurs in, is the number of components required in the model. However, models often have similar PRESS even when fewer components are being used. Adding additional components could lead to only marginally lower PRESS, not adding much to the model for the additional complexity and perhaps only explaining additional random noise. Therefore the SAS® manual recommends a statistical randomisation based test developed by van der Voet (1994), which selects the model as the smallest number of components with a PRESS not significantly larger than the minimum PRESS. The null model and alternative models are defined as: H_0 - The squared

residuals from the model with and without the extra component have the same distribution and H_1 - The squared residuals from the model with and without the extra component have a different distribution. Monte Carlo simulations are then used to determine whether the observed distributions from each model (with or without the extra component) are likely to have occurred by chance alone. A p-value for the addition of each component is provided. The SAS® manual recommends the number of components as the first component with a corresponding van der Voet's test p-value greater than 0.1 (SAS, 2008, p. 4783), hence H_0 is not rejected (at the 10% alpha level). In this instance it is concluded that the squared residuals from the model with and without the extra component, could have the same distribution and subsequently the extra component is not required in the model.

Another option to select the number of components is to use the calculation of R^2-CV initially defined in 4.2.1.1 (González et al., 2011, Le Cao et al., 2008). R^2-CV (sometimes called Q^2) is calculated as 1- (PRESS/SS). Where PRESS is the residual sum of squares from the predicted model (as above) and SS is the total variation observed in the Y variable. PRESS/SS equates to the proportion of Y variation the model is not explaining, hence 1-(PRESS/SS) describes the amount of variation the model is explaining. R^2-CV can be calculated for each component representing the amount of variation each additional component is adding to the model, or it can be presented cumulatively (the total amount of variation the model is explaining). Historically predictive significance was determined using a 5% alpha level and the square root of the mean squared error of prediction (RMSEP). As it has been square rooted, the unit is based on the STD from the model not the variance. RMSEP=$\sqrt{(PRESS/N)}$. As R^2-CV uses the variation unit, to retain the same significance level, a criteria of $1\text{-}0.95^2 = 0.0975$ was traditionally used as a cut off for the R^2-CV (Eriksson et al., 2006a, Le Cao et al., 2008). Therefore, it was decided that if the R^2-CV for the addition of any component is <0.0975 then it will indicate that component is not required.

Both of these conditions will be examined and used to determine the number of components as summarised in the flow diagram in section 4.2.6.

4.2.1.4. Number of variables to extract

In SPLS, the final model does not contain all of the original variables. Therefore a decision has to be made regarding how many variables to keep (extract) for the final model, which are hopefully truly predictive of the signal and not merely predicting noise in the data (over fitting). The mixOmics function version 3.0 uses the loading vector (which indicates the size of the contribution of each variable to the prediction model) to order the variables. For each fold and run, this order can be used to reflect the order of importance of each variable.

The R^2 of the model will continue to rise as more and more variables are fitted in the model until there is perfect fit of the data. To avoid over fitting, González et al. (2011) recommend calculating the R^2-CV and plotting it against the number of variables in the model. The point at which adding an extra variable makes the prediction under CV worse, suggests that no more variables are required in the model. Many authors recommend the use of CV to determine the optimum number of variables to retain for other penalised modelling methods (Abraham et al., 2013, Alexander and Lange, 2011, Li and Sillanpaa, 2012, Varma and Simon, 2006, Abraham et al., 2012). In addition, Le Cao et al. (2011) use this approach with SPLS. One particular concern is that three authors have recently found that the models tend to include more variables than the optimum (Li and Sillanpaa,

2012, Ayers and Cordell, 2010, Abraham et al., 2013). In addition, Varma and Simon (2006) identified that the predictive performance which would be observed on an independent test set is overestimated by CV.

Based on the guidance by González et al. (2011), it was decided to determine the number of variables to include in the final model by repeatedly running a CV model and extracting a different number of variables each time (from 1, 2, 3, 4, 5, 10, 15 and up to 250 variables). The R^2-CV of each model will be calculated and plotted against the number of variables that model extracted. The number of variables with the maximum R^2-CV will be selected as the optimum number of variables to extract (unless there was no clear optimum, then a decision is made based on the approximate highest R^2-CV with possibly a smaller number of variables extracted).

4.2.1.5. Further model considerations (number of folds, runs and model selection)

Another key decision to make when using CV is how many folds to use. When modelling the 'all subjects' dataset (N=912), folding the data into 10 folds, results in 91 (or 92) subjects being predicted each fold. This seems a large enough number to use. Running the SPLS through once, an 'unlucky' split of patients may by chance make the selected model dependant on the way the subjects were randomly assigned. To get a robust model which is the same no matter which fold patients are assigned to, it was decided to repeat the 10-fold CV a number of times (runs). The number of runs was explored using 50 to 200 and results indicated that 50 was able to produce a stable model with a feasible running time (A comparison of running times is shown in section 10.3). After completing this work, it was observed that as few as three runs had been used in the literature (Long et al., 2011), this decision is reassessed on the 'GWAS SNPs' dataset in section 6.4.2, when running times of the model were substantially increased.

Fifty runs of 10-fold CV will produce 500 models with different variables extracted. To decide upon a final model, variables need to be identified which are consistently being chosen irrespective of the folds used. After 50 runs of 10-fold CV, a dataset is created containing the name of each variable and the number of times it was selected in the top set of variables extracted. This approach is used by Magidson (2011) and Gonzalez et al. (2011) who in data examples which were not using SNP data, only select variables to be kept in the final model if they are chosen in all folds and runs. A stable model is required where variable selection is not dependent on how the data is split into folds. To ensure this and to avoid SNPs that are inconsistently selected, it was decided to follow Magidson (2011) and Gonzalez et al. (2011) insisting that a variable should be selected by all 50 runs. However, because of modelling SNP data, there is a chance that a SNP is not selected in a fold due to insufficient variation (i.e. all of the genotypes of 1 or 2 are in the test set). If such a case occurs then that SNP cannot be selected in that fold as being predictive of RA severity however may be selected in the other 9/10 folds. To try to accommodate for rare but important SNPs, it was decided to relax the rule of having to be in 100% of the folds to only requiring SNPs to be selected in 80% instead.

Whilst different rules may be required for analysis of the 'GWAS SNPs' dataset, it was felt that with 10 folds, retaining variables which were selected 80% of the time (8/10 folds) in all of the 50 runs, would retain key predictor variables for the final model. This method was entitled the 'percentage fold' method. Stricter criteria resulted in rarer SNPs not having a chance of being selected as

potentially they didn't have enough variation to be modelled in all of the folds. More generous criteria resulted in too many variables for the final model which could result in an over fitted model.

4.2.2. Data description and imputation

Using the 'all subjects' dataset (N=912), SPLS methods will be used to select the most important variables predictive of the Larsen score from 368 SNPs and 19 environmental variables. Section 2.2.3 provides an explanation of this subset of data and Appendix C contains a full list of SNPs including which chromosomes they are located on. The environmental variables defined below are chosen because they are measurable at disease diagnosis and though in this study the data was collected at recruitment, this behaviour should not be confounded with disease severity over time.

- Alcohol use: Do you drink alcohol? (Yes/No),
- Alcohol quantity: During the past 30 days, on how many days did you have at least one drink of alcohol? (None, <5 days, 5-10 days, more than 10 days),
- Categories of the shared epitope as defined by Tezenas du Montcel et al. (2005) S1, S2, S3d, S3p and X and Number of copies of the shared epitope (0,1,2),
- Sex (M/F),
- Age, age at time of disease diagnosis, age at onset of symptoms,
- Disease duration & symptom duration,
- Smoking status (never, former, ever), average number of cigarettes per day, Smoking duration (years), time since quitting smoking and number of smoking pack years (calculated as number of years smoking * average number of cigarettes per day / 20).

As described in section 3.8, missing data is required to be imputed for PLS modelling. This can be performed within the model fitting process however it was decided to impute all missing data prior to model fitting for two reasons:

1. If missing data is imputed during the model fitting and multiple runs of multiple-fold CV are performed, then the model would be doing the imputation multiple times and hence increase the time it takes to run the model.
2. Using the mixOmics version 3.0 function to perform SPLS CV, each time the model is fitted on the M-1 folds, a different set of variables can be chosen. These variables are then used to predict the group of subjects in the Mth (left out) set. If those subjects have any of the chosen variables with missing data, then a prediction for that subject cannot be made. Therefore the imputation has to be performed prior to the modelling to enable the CV prediction to be calculated for every subject no matter which variables are selected.

Whilst the environmental variables recorded are relatively complete as shown in section 2.2.2, all subjects have a least one missing SNP when using the 'GWAS SNPs' dataset. Therefore to apply SPLS, imputation methods are required to be applied prior to the modelling.

Although there are many methods using LD to impute missing SNP data, the data would still require imputation of the environmental variables and imputation of SNPs when LD imputation methods were not possible. Therefore, SPLS models created using two methods of imputation are explored for the 'all subjects' dataset; 'quick' imputation and NIPALS imputation as described below. No LD imputation methods are investigated at this stage as the 'all subjects' dataset consists of 368 SNPs

spread through the genome and hence were not in strong LD. The effect of using LD to impute missing SNP data is further investigated using the 'GWAS SNPs' dataset in chapter 5.

'Quick' imputation

'Quick' imputation is defined as imputing the environmental continuous variables with the mean and environmental categorical variables with the mode. The HLA-DRB1 (S1, S2, S3p, S3d and X) variables are imputed with their modal value similar to the environmental variables because there are only five variables and computationally it would not take long to impute with the mode. However, all SNPs are imputed with a 0 making them the wild type homozygous (most frequently occurring homozygous genotype). This method is much quicker to run on 'GWAS SNPs' dataset as all missing SNP data is replaced with a zero rather than having to count the most frequently occurring genotype. In addition, it may provide a contrast to the value given using the NIPALS algorithm described below which uses correlation between the SNPs.

Non-linear estimation by iterative partial least squares (NIPALS)

NIPALS imputes missing data iteratively through multiple bivariate regression models as described for PCA analysis in section 3.5. . The slope for each variable in X is iteratively estimated using simple bivariate least squares regression lines using one variable as the y-variable and another variable as the x-variable. Therefore the correlations between all of the variables in the same matrix are used to estimate the missing data based on the location that subject has in relation to other subjects with similar data in other variables (Eriksson et al., 2006a, p. 65).

Using the function NIPALS in R, the missing values are replaced by the estimated values from the reconstituted matrix derived iteratively through the multiple bivariate regressions.

4.2.3. Effect of missing data on the model fitting

Many authors warn that the PLS algorithm with NIPALS imputation can cause the models to select variables with large amounts of missing data and has a loss of robustness as the missing data approaches 20% (Rannar et al., 1995, Trygg and Wold, 2002, Pedreschi et al., 2008). Pedreschi et al. (2008) observed NIPALS imputation to cause an artificial reduction in the variance. The more data that is imputed, the more the variance is artificially reduced. Hence, the more data a variable has imputed, the more likely it is to be selected in the model. Trygg and Wold (2002) therefore recommend keeping the imputed data for any variable to be less than 20%. They suggest that the artificial reduction in variation for variables with less than 20% imputed has little influence on the variable selection for the model.

This could be problematic for the GoRA data, as although the environmental data is quite complete, some SNPs have at least 40% missing data. The reason for this missing data is that although the original call rate may have been acceptable, data has been merged together from different sources which are measured on different numbers of patients (section 2.2.3). For example, a SNP measured by the Sheffield group on 400 of the GoRA subjects may have a 99% call rate, however when merged with the full GoRA set of patients (912 in the 'all subjects' dataset), then the SNP has 56% of subjects missing.

To determine if this was a concern in this analysis, using NIPALS imputation, SPLS models were fitted and the chosen variables extracted. The amount of missing data in the full dataset was then compared with the amount of missing data in the variables chosen by the model. Including all environmental variables and SNPs in the 'all subjects' dataset, each variable on average had 8.0% of data missing (8.21% for the SNPs alone). In comparison, the variables selected in the NIPALS Larsen score prediction model had on average 17.6% missing data (20.6% for the SNPs alone). It was therefore true that the selected variables had more missing data than the variables not selected. In fact, 19 of the 25 SNPS with missing data greater than 20% were selected for the final model. It was therefore decided to remove any SNPs with >20% of their data missing. The 'quick' imputation models were also investigated and as they selected similar variables to the NIPALS model, this approach was applied for all models. The 'quick' imputation method imputed missing SNPs with a 0 and hence would also suffer from artificially reducing the variation.

4.2.4. Transformations to the Larsen score prior to modelling

Eriksson et al. (2006a) warn that PLS works best when the data is normally distributed. Due to the shape of the Larsen score distribution (Figure 2.1) and the wide varying disease duration in this sample (Figure 2.2), adjusting the Larsen score for disease duration before modelling and modelling the natural logarithm of the data will be investigated. Models investigating Larsen score alone and Larsen score/disease duration (both unlogged, logged and with subjects having zero Larsen scores excluded) were examined to see whether the predictions and performance of the model were consistent. Modelling the Larsen score alone, (with either imputation method), performed better in terms of R^2-CV and correlation between actual and predicted Larsen score than any investigated transformation (Table 4.1).

One reason for the Larsen score performing better than the Larsen score/disease duration could be that the Larsen score progression rate is unlikely to be constant over time. Suppose the Larsen score progressed as in Figure 4.1, a subject with 30-years of disease duration, would have their Larsen score adjusted much more than a subject with 15-years of disease duration. However, their Larsen scores could be similar, having levelled out over this time. Therefore, adjusting for disease duration in a linear way may be inappropriate.

Table 4.1 Summary of models fitted using various transformations and imputations

Model	N[a]	N variables asked to be extracted	N (%) selected for Final model	Final model R^2-CV[b]	r corr-elation[c]	% Xvar[d]	% Y var[e]
Larsen ('Quick' imputation)	1	65	39 (60%)	0.561	0.577	11.1	33.3
Larsen (NIPALS)	1	100	64 (64%)	0.572	0.592	6.5	35.0
Larsen /disease duration (NIPALS)	2	40	33 (82.5%)	0.36	0.433	15.0	18.7
Log (Larsen/disease duration) (NIPALS)	2	50	44 (88%)	0.424	0.484	14.7	19.6
Log (Larsen /disease duration) no 0 (NIPALS)	1	65	24 (36.9%)	0.315	0.349	10.53	12.2

a: Number of components selected to be extracted from the model

b: The proportion of variance in the response explained by the regression model when the chosen variables are refitted on the data using 10-fold CV

c: Re-fitting the chosen variable on all of the data to retrieve the parameter estimates and then fitting this model on all patients (no CV) this is the correlation between the actual and predicted Y

d/e: Percentage of X/Y variation explained by the model.

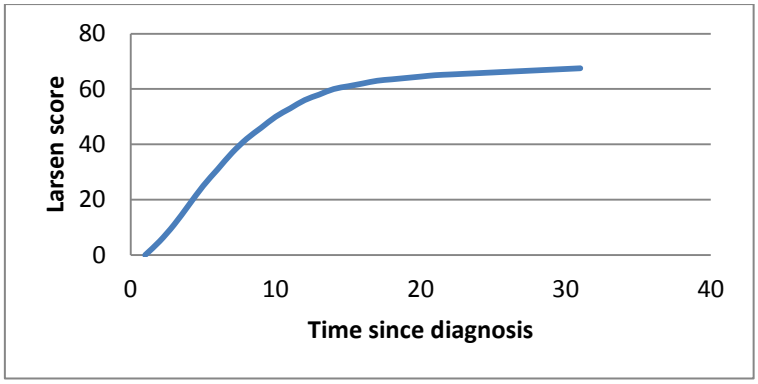


Figure 4.1 Example of possible Larsen score progression rate

4.2.5. Univariate modelling and estimation of the effect size

In many applications, bootstrapping is used to calculate 95% confidence intervals around the parameter estimates. However, it is not appropriate in PLS analysis because the beta coefficients for each variable would be 'weakened' if variables in the model are correlated. For example, if two variables are in complete correlation (r=1), then the size of their beta coefficient will be halved. This may cause important variables to have smaller beta coefficients in the model and hence 95% confidence intervals which overlap zero. Therefore, in PLS it is not appropriate to interpret the size of the coefficient without taking into account the correlation of variables in the model. Instead, it was decided to quantify the direction & size of effect by presenting the median Larsen score for each genotype and testing each SNP univariately. When analysing the 'all subjects' dataset, in order to suitably model the inflated number of subjects with a Larsen score equal to zero (Figure 2.1), a ZINB model was found to fit the data better than a zero inflated Poisson model, negative binomial (NB) model or Poisson model.

In order to adjust the univariate ZINB modelling for each SNP to predict the Larsen score, it is important to adjust for potential confounders such as disease duration. Literature of modelling strategies recommend using medical knowledge to select important variables rather than stepwise, forward or backward selection methods (Harrell, 2001, p58). Medical knowledge has been used to reduce the many environmental variables to the few selected for the multivariate modelling as listed in section 4.2. However, this list still includes many highly correlated variables which would unnecessarily complicate a univariate model whose objective is to assess the importance of each SNP.

A simple forward selection strategy will therefore be applied to create a standard list of environmental factors which will be adjusted for, when analysing the importance of SNPs. All factors are assumed to have a linear relationship with Larsen score and are included in both the count and zero inflation part of the model. Likelihood ratio (LR) tests are used to compare nested ZINB models (-2*Log Likelihood of fewer variable model + 2 * Log likelihood of greater variable model) and assessed using the chi-square distribution on two degrees of freedom.

The defined strategy begins with no variables/factors in the model (the base model). Each environmental factor is added one at a time and the most significant LR test indicates the 1st factor to include. The base model is then amended to include the 1st important factor and the remaining factors are again added one at a time. The factor with the most significant LR test is then added

into the base model and the process repeated until none of the LR tests are significant at the 5% alpha level.

In section 4.2, BMI and ACPA were not included as potential environmental predictors due to them being measured post diagnosis of disease and hence, their value could be influenced by disease severity. However, it was decided to include them in the univariate analysis as they were included in the 'GWAS SNPs' dataset multivariate modelling in section 6.2.2.

Using the above strategy, the first parameter to enter the model was the disease duration ($P<1x10^{-15}$). The ZINB model including disease duration was then fitted and compared against fitting disease duration plus each remaining factor one at a time. The second variable to enter the model was ACPA category (positive/negative) ($p<1x10^{-15}$). Continuing the process, the third variable to be entered was symptom duration (p=0.00015), followed by BMI (p=0.000769), age at onset of symptoms (p=0.006617) and ACPA value (p=0.005992). At this point no further variables were significant at the 5% level and it was decided not to enter any further variables. The 5% alpha level was selected so that only strongly predictive variables were included and the model was not over fitted. Therefore, the environmental variables to include in all ZINB models to investigate SNPs will be Larsen score = Disease duration + ACPA category + Symptom duration + BMI + age at onset of symptoms and ACPA value. The significance of these variables when fitted together in the ZINB model is shown in Table 4.2.

Table 4.2 Base model for testing SNPs using ZINB models

Count model coefficients (NB with log link):				
	Estimate	STD	z value	Pr(>|z|)
(Intercept)	3.370543	0.208103	16.197	$< 2 \times 10^{-16}$
Disease duration	0.026751	0.0060851	4.396	1.1×10^{-5}
ACPA category	0.201284	0.0877474	2.294	0.0228
Symptom duration	0.010815	0.0059292	1.824	0.0682
BMI	-0.01693	0.005488	-3.085	0.0020
Age at onset of symptoms	-0.00513	0.0022185	-2.314	0.0206
ACPA value	0.002086	0.0008064	2.586	0.0097
Log(theta)	0.645284	0.0522094	12.36	$< 2 \times 10^{-16}$
Zero-inflation model coefficients (binomial with logit link):				
	Estimate	STD	z value	Pr(>|z|)
(Intercept)	-2.14796	0.85796	-2.504	0.0123
Disease duration	-0.00446	0.038703	-0.115	0.9082
ACPA category	-1.12192	0.333487	-3.364	0.0008
Symptom duration	-0.07116	0.037301	-1.908	0.0564
BMI	0.049705	0.020311	2.447	0.0144
Age at onset of symptoms	0.020651	0.009057	2.28	0.0226
ACPA value	-0.00736	0.003848	-1.913	0.0558

The above model will be used as the base model and each SNP will be entered into the count and zero-inflation part of the model. A LR test as described above, will be calculated and compared against the chi-square distribution on two degrees of freedom, to determine whether the SNP has a significant contribution to the model. In the rare case that all subjects with a Larsen score equal to zero have the most frequent homozygous genotype, hence there is no variation in the zero-inflation part of the model for that SNP, the SNP will only be entered into the count part of the model and tested against the chi-square distribution on one degree of freedom.

When the 'GWAS SNPs' dataset is being investigated, there is no longer an inflation in the number of zeros (Figure 6.1), hence it is more appropriate to use the NB distribution. The Poisson distribution was also considered, however, a LR test indicated the NB model was more appropriate, suggesting over dispersion relative to a Poisson model. The same environmental variables will be included in the model as defined in Table 4.2 as it may be biased to reassess the covariates based on a subset of patient's data. The p-value corresponding to the SNP in the model will be presented.

A univariate one SNP at a time model may over exaggerate the size of the SNP effect, as it is examined in isolation to the rest of the model. However, it will give indication of whether the SNPs would have been identified as statistically significantly important if they were analysed in a univariate way. This provides an alternative view on whether the model is selecting important SNPs. In addition, the medians will show the average difference in Larsen score across the genotypes.

4.2.6. Summary of SPLS model creation process 'percentage fold' method

Using guidance from many authors (González et al., 2011, Le Cao et al., 2008, Eriksson et al., 2006a, SAS, 2008, Long et al., 2011) the following flow diagram was created to detail the model fitting process entitled 'percentage fold' method used on the 'all subjects' dataset throughout section 4 (Figure 4.2).

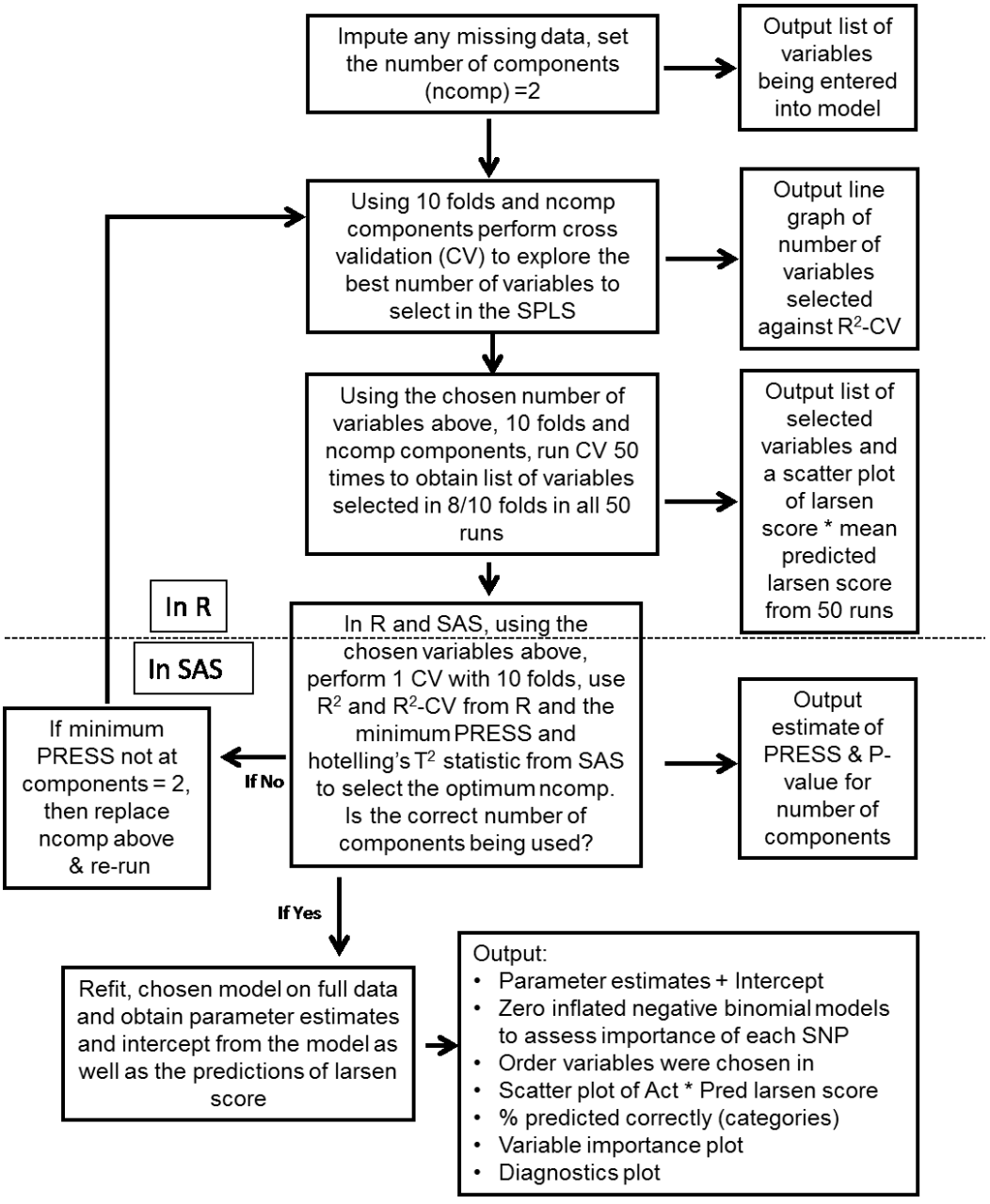


Figure 4.2 Iterative procedure for fitting PLS models using the 'percentage fold' method

4.3. 'Percentage fold' method results when using 'quick' imputation

SPLS models using the 'percentage fold' method (section 4.2.6) were fitted using the 'all subjects' dataset. The data consisted of 912 subjects using 387 X variables (368 SNPS, 19 environmental variables) after applying the 'quick' imputation method described in 4.2.2 (excluding SNPs with >20% missing data per 4.2.3). Using a two component model, the gain in R^2 (not under CV) by adding the second component was just 0.0146 (0.5864 with two components and 0.5718 with one component). For the second component, R^2-CV =0.016 suggesting that only one component was required, as the R^2-CV was less than 0.0975. Using the van der Voet's test, a two component model had the smallest PRESS with p-value > 0.1. However, as the model containing just the 1st component had a p-value of 0.0960, this indicated it was only just not being selected as the optimum number of components. After investigating the amount of variation the additional component was able to explain (36.3% of the Y variation with two components compared to 33.8% of the Y variation with one component), it was decided more appropriate to use the simpler model with one component. Therefore all models below are using a one component model.

4.3.1. Number of variables to extract ('quick' imputation)

Following the selection of the number of components, the number of variables to extract was determined by running one SPLS model at a time extracting from one to 250 variables as described in 4.2.1.4. The maximum R^2-CV was observed at 95 variables (R^2-CV =0.599, Figure 4.3).

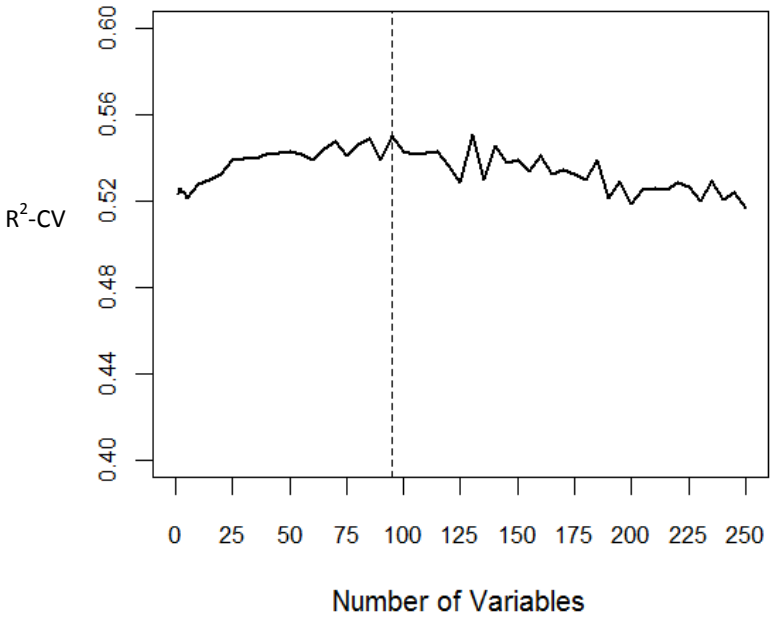


Figure 4.3 Plot of the R^2-CV versus the number of selected variables ('quick' imputation)

4.3.2. Selection of the final model ('quick' imputation)

The model was run 50 times using 10-fold CV and extracting 95 variables in each fold and each run. To enable the variables to be sorted into an order of importance in each fold and run, variables were ranked (most predictive to least predictive) according to the amount of Larsen score variation explained (based on the values of each variables loading in the loading vector). Any variables not in the top 95 were given a loading of 0 (and ranked equal last) so as to not include them in the model. 58 variables were selected in 8/10 folds, in all 50 runs.

To obtain an overall order of importance, the median rank was calculated first across the 10 folds within each run, and then across the 50 runs and is presented in the rank column in Table 4.3.

Using the chosen model, CV was performed resulting in an R^2-CV of 0.575. Table 4.3 reveals the final model, along with the median rank order, the number of subjects and median Larsen score for each genotype and the ZINB LR p value for SNPs (as described in section 4.2.5). Only 13 of the 48 SNPs (27%) selected were significant at the 5% alpha level. Although some of the highest ranked SNPs (rs2075800 ranked 5 and rs2844479 ranked 6) have univariate p-values as low as $<1.0 \times 10^{-15}$, the SNP ranked 7th rs2071592, has a univariate p-value of 0.453. Investigation of the median Larsen score alone for each genotype (0,1,2) could not explain this large difference in p-values, as in some cases, a clear increase or decrease in median Larsen score linearly across the genotypes was found not to be statistically significant.

As the p-value tells us how likely our null hypothesis is to be true based on the strength of the evidence in the observed data, it appears heavily reliant on the numbers of subjects and variation in Larsen score within each genotype for both the zero part and count part of the ZINB model. For example, for four non-significant SNPs (rs2071592, rs5029937, rs1800629 and rs7752903), all subjects with a Larsen score of zero had the most frequent homozygous genotype (0) and hence there is zero variation in the zero inflation part of the model. Therefore, the SNP could only be fitted in the count part of the ZINB model and the LR p-value (calculated on one degree of freedom) was found to be non-significant. Of particular concern, is the large change in significance when rs2071592 is analysed in section 4.4.2. After NIPALS imputation, at least one non-zero genotype now has a Larsen score of 0. With the new present variation, the p-value becomes highly significant $p=9.24 \times 10^{-11}$. Of the other SNPs, rs1800629 is not selected at all in the NIPALS model, however, rs5029937 and rs7752903 are both significant at the 1% alpha level (p=0.0019 and p=0.0063 respectively). This appears to indicate a problem with the use of univariate ZINB modelling of small data samples or low frequencies of genotypes.

In the multivariate PLS analysis, variables are selected according to the size of their loading value. This can be thought of as the cosine of the angle between the original variable and the line of best fit through the data rotated to have most correlation with the Y data. Variables with high loadings are hence selected for inclusion in the model. This is a very different way of assessing importance of variable prediction compared to the univariate p-value setting. Given the instability of the p-value shown above, the multivariate method is considered less influenced by small changes in the data caused by the method of imputation.

Of particular interest, the shared epitope variable (coded as 0 RAA motif, 1 RAA motif or 2 RAA motif copies), was not selected in the SPLS model. However, the DRB1 S1 and S2 shared epitope

variables both appeared and were consistently in the direction documented in previous literature. This added supportive evidence that for severity prediction, using the Tezenas du Montcel et al. (2005) DRB1 coding is more appropriate than counting the alleles with the RAA shared epitope.

The proportion of missing data in the selected variables was calculated in order to verify that preference was not being given in the PLS selection process to variables with more missing data (section 4.2.3). The average missing data in all the variables available for model selection was 5.17% whereas the mean amount of missing data in the variables which were selected was only slightly higher at 6.59%. Therefore, removing SNPs with >20% missing data has prevented variables with a higher frequency of missing observations being selected more frequently.

Table 4.3 Final model when using 'quick' imputation

Variable name	Rank	Genotype (0, 1 or 2): N subjects: Median Larsen score			ZINB LR p value	Description
Disease duration	1					Increase in Larsen score for increase in disease duration
Symptom duration	2					Increase in Larsen score for increase in symptom duration
Age at onset of symptoms	3					Decrease in Larsen score for increase in age at onset of symptoms
Age at time of diagnosis	4					Decrease in Larsen score for increase in age at disease duration
rs2075800	5	0: 358: 17	1: 410: 34	2: 144: 32.5	<1.0E-15	C/T variant located on chromosome 6 at 31777946. Missense function in HSPA1L.
rs2844479	6	0: 393: 18	1: 391: 32	2: 128: 38	<1.0E-15	G/T variant located on chromosome 6 at 31572956. Unknown function.
rs2071592	7	0: 425: 20	1: 385: 32	2: 102: 33	0.453	A/T variant located on human chromosome 6 at 31515340. Intron variant of NFKBIL1
rs2242653	9.5	0: 647: 26	1: 237: 32	2: 28: 44.5	4.74E-12	C/T variant location on human chromosome 6 at 31676015. Missense variant in LY6G6F
rs9366826	10	0: 560: 24.5	1: 316: 31	2: 36: 31	0.0039	C/G variant located on chromosome 6 at 33619184. Intron function in Inositol 1,4,5,-trisphosphate receptor, type 3 (ITPR3)
rs805292	10.5	0: 651: 25	1: 233: 34	2: 28: 34	5.83E-11	C/T variant located on chromosome 6 at 31690259. 8167 bases upstream of ABHD16A, 4428 bases downstream of LY6G6F, 8167 bases upstream of LY6G6E, 4428 bases downstream LY6G6D, 498 bases upstream of LY6G6D, 1112 bases upstream of C6ORF25, 4808 bases downstream of DDAH2 and 8349 bases downstream of CLIC1.
rs394581	11.5	0: 523: 32	1: 335: 20	2: 54: 29	0.196	C/T variant located on chromosome 6 at 159482771. 5' untranslated region of the TAGAP gene. Associated with the risk of Rheumatoid Arthritis
rs443198	12.5	0: 413: 22	1: 382: 31.5	2: 117: 37	<1.0E-15	C/T variant located on chromosome 6 at 32190656. Synonymous variant in NOTCH4 associated with activity in RA synovium
rs26232	13.5	0: 491: 29	1: 331: 27	2: 90: 16	0.0071	C/T variant located on chromosome 5 at 102596720. Intron variant in C5orf30 associated with RA susceptibility and severity
Alcohol quantity	14					Decrease in Larsen score with increase in alcohol quantity (coded as 0=None, 1=<5 days, 2=5-10 days, 3=more than 10 days)
rs182429	14	0: 366: 31.5	1: 430: 26	2: 116: 21.5	0.164	C/T variant located on chromosome 6 at 159469574. 3390 bases upstream of TAGAP (Eyre et al., 2010b)
rs2256965	17.5	0: 463: 24	1: 354: 31	2: 95: 33	<1.0E-15	C/T variant located on chromosome 6 at 31555130. Intron variant of LST1: Leukocyte-specific transcript 1 protein
rs2872507	19	0: 321: 23	1: 387: 27	2: 204: 32.5	0.118	C/T variant located on chromosome 17 at 38040763. Intergenic, 6614 bases downstream of ZPBP2.
Alcohol use	19.5					Decrease in Larsen score with using alcohol vs not using alcohol
rs220704	19.5	0: 704: 26	1: 196: 37.5	2: 12: 27.5	0.159	A/T variant located on chromosome 6 at 46865758. Intron variant in GPR116.
rs4535211	19.5	0: 303: 33	1: 420: 26	2: 189: 20	0.032	A/G variant located on chromosome 3 at 17072997. Intron variant in PLCL2 (Bowes et al., 2012)
rs2431697	22.25	0: 337: 31	1: 415: 27	2: 160: 19.5	0.297	C/T variant located on chromosome 5 at 159879978. Intergenic region, between PTTG1 and microRNA miR-146a associated with Lupus & psoriasis
Smoking pack years	25					Decrease in Larsen score with increase in amount of smoking pack years

Variable name	Rank	Genotype (0, 1 or 2): N subjects: Median Larsen score			ZINB LR p value	Description
drb1nos2	26	0: 483: 26	1: 359: 29	2: 70: 39.5	0.982	Increase in Larsen score with presence of K-R-A-A shared epitope sequence
rs5029937	28	0: 830: 27	1: 79: 37	2: 3: 31	0.940	G/T variant located on chromosome 6 at 138195151. Intron variant in TNFAIP3 associated with RA risk (Orozco et al., 2009)
drb1nos1	31	0: 659: 29	1: 235: 26	2: 18: 10	0.0621	Decrease in Larsen score with presence of A-R-A-A or E-R-A-A
rs10917214	32	0: 331: 30	1: 421: 26	2: 160: 25.5	0.127	A/G variant located on chromosome 1 at 22652501. Intergenic with unknown function.
rs932744	32.75	0: 374: 24	1: 412: 28.5	2: 126: 36.5	0.100	C/G variant located on chromosome 6 at 150390663. 380 bases upstream of ULBP3
rs2009345	33	0: 365: 24	1: 422: 28	2: 125: 37	0.124	A/G variant located on chromosome 6 at 150389748. Intron variant in ULBP3.
rs1800629	35.25	0: 648: 26	1: 237: 29	2: 27: 43	0.454	A/G variant located on chromosome 6 at 31543031. 313 bases upstream of TNF, 931 bases upstream of LTA and 5305 bases downstream of LTB.
rs1076933	35.5	0: 354: 29	1: 410: 28	2: 148: 21.5	0.159	A/G variant located on chromosome 22 at 45198494. Intron variant in ARHGAP8 and PRR5-ARHGAP8
Smoking duration	35.5					Larsen score is decreased by an increase in duration the subject has smoked.
rs7752903	37.5	0: 844: 27	1: 66: 37.5	2: 2: 31	0.965	G/T variant located on chromosome 6 at 138227364. Intergenic with unknown function. In a 109 kb DNA segment that spans the TNFAIP3 gene possibly related to SLE
rs228975	38.25	0: 382: 22.5	1: 404: 28	2: 126: 35	0.225	C/T variant located on chromosome 22 at 37542201. Intron variant in IL2RB.
rs4777183	38.5	0: 300: 31	1: 418: 27	2: 194: 20	0.482	C/T variant located on chromosome 15 at 70004775. Intergenic with unknown function.
rs5029938	39	0: 852: 28	1: 60: 18.5	2: NA: NA	0.484	C/T variant located on chromosome 6 at 138195633. Intron variant in TNFAIP3 associated with RA risk (Orozco et al., 2009)
rs5029939	41	0: 828: 27	1: 82: 35.5	2: 2: 31	0.045	C/G variant located on chromosome 6 at 138195723. Intron variant in the TNFAIP3 associated with RA risk (Orozco et al., 2009)
rs4265819	42	0: 753: 26	1: 151: 36	2: 8: 29.5	0.007	A/G variant located on chromosome 16 at 62617217. Intergenic with unknown function
rs2327832	43	0: 507: 25	1: 351: 32	2: 54: 24	0.558	A/G variant located on chromosome 6 at 137973068. Intergenic with unknown function.
rs2476601	44.75	0: 647: 26	1: 233: 33	2: 32: 44	0.114	A/G variant located on chromosome 1 at 114377568. Missense variant in PTPN22 associated with RA risk and possibly ACPA positivity.
rs9565072	45	0: 452: 30	1: 363: 26	2: 97: 17	0.143	C/T variant located on chromosome 13 at 74639799. Intron variant in KLF12 which was investigated for risk of RA (Eyre et al., 2010a) however no associated found.
rs864745	45.5	0: 312: 31	1: 402: 27	2: 198: 26.5	0.283	A/G variant located on chromosome 7 at 28180556. Intron variant in JAZF1 which has previously been associated with Type II diabetes
rs13061519	45.75	0: 854: 28	1: 56: 19.5	2: 2: 40	0.266	C/T variant located on chromosome 3 at 173235402. Intron variant in NLGN1.
rs2027276	46	0: 852: 28	1: 60: 19.5	2: NA: NA	0.552	G/T variant located on chromosome 6 at 138174328. Intron variant in LOC100130476.
rs2230926	46.75	0: 827: 27	1: 83: 36	2: 2: 31	0.040	G/T variant located on chromosome 6 at 138196066. Missense variant in TNFAIP3 associated with RA risk (Orozco et al., 2009)
rs8045689	47	0: 492: 26	1: 342: 30.5	2: 78: 31.5	0.111	C/T variant located on chromosome 16 at 28988269. Intron variant in SPNS1.
rs5980742	47.5	0: 393: 32	1: 299: 26	2: 220: 25	0.062	G/T variant located on chromosome X at 70321631. Intron variant in FOX04.
rs10919563	48.5	0: 732: 28.5	1: 171: 26	2: 9: 21	0.146	A/G variant located on chromosome 1 at 198700442. Intron variant in PTPRC
rs12403075	49.5	0: 723: 29	1: 180: 25.5	2: 9: 21	0.254	C/T variant located on chromosome 1 at 198721192. Intron variant in PTPRC
rs6473517	50.5	0: 703: 29	1: 196: 21.5	2: 13: 18	0.084	G/T variant located on chromosome 8 at 84857572. Intergenic with unknown function
rs26510	51	0: 447: 31	1: 371: 23	2: 94: 28	0.591	C/T variant located on chromosome 5 at 96125910. Intron variant in ERAP1
rs6000570	51	0: 211: 26	1: 456: 26	2: 245: 33	0.575	A/G variant located on chromosome 22 at 37514339. 8736 bases upstream of TMPRSS6 and 7541 bases downstream of IL2RB.
rs6933404	52.5	0: 508: 25	1: 351: 32	2: 53: 24	0.698	C/T variant located on chromosome 6 at 137959235. Intergenic with unknown function
rs7722135	53.25	0: 582: 26	1: 281: 32	2: 49: 36	0.299	C/T variant located on chromosome 5 at 86294669. Intergenic with unknown function.

Variable name	Rank	Genotype (0, 1 or 2): N subjects: Median Larsen score			ZINB LR p value	Description
rs6932056	53.25	0: 837: 27	1: 73: 36	2: 2: 31	0.032	C/T variant located on chromosome 6 at 138242437. Intergenic with unknown function.
rs11586238	59	0: 548: 30	1: 310: 21	2: 54: 28.5	0.239	C/G variant located on chromosome 1 at 117263138. Intergenic with unknown function however possibly associated with RA risk (Raychaudhuri et al., 2009)
rs6927172	59.5	0: 529: 26	1: 332: 32	2: 51: 24	0.545	C/G variant located on chromosome 6 at 138002175. Intergenic with unknown function.
rs10499197	60.25	0: 846: 27	1: 64: 36.5	2: 2: 31	0.081	C/T variant located on chromosome 6 at 138132516. Intergenic with unknown function.
rs13242262	60.25	0: 414: 30.5	1: 379: 27	2: 119: 19	0.401	A/T variant located on chromosome 7 at 128591364. 1276 bases downstream of IRF5 (associated with SLE risk), 2870 bases downstream of TNP03.

4.3.3. Assessing the predictive ability of the model ('quick' imputation)

The predictive ability of the model was assessed by re-fitting it on the patients used to form the model, not using CV (Figure 4.4). The model achieved a correlation between the actual Larsen score and the predicted Larsen score of r=0.599.

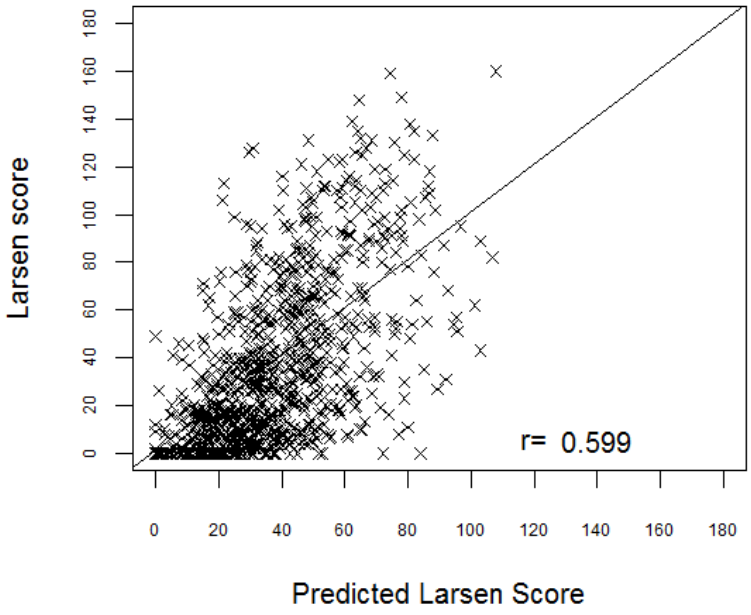


Figure 4.4 Scatter plot of actual vs. predicted Larsen score ('quick' imputation)

The mean absolute difference between the actual and predicted Larsen score was 21.7 (Median = 17.7, STD =17.6, Minimum=0, Maximum=97.2). As shown by the minimum and maximum, there was a wide spread in the prediction error with some subjects very poorly predicted (up to 97.2 Larsen score points from the true Larsen score) and some subjects predicted correctly.

4.3.4. Variance partitioning ('quick' imputation)

This research aim is to predict RA severity at disease diagnosis. Therefore any model should only include variables measurable pre-diagnosis. At such a time the disease duration would be zero and it is likely symptom duration would also be very low due to the rapid detection and confirmation of RA in modern hospitals. It is therefore of importance to remove the variation in the Larsen score model which is attributable to the disease duration and symptom duration. It is useful to quantify how much of the Larsen score variability is explainable by the environmental and genetic variables

alone. There are a number of ways the variation in the Larsen score, which the final model is able to explain can be explored. Unfortunately, the R package mixOmics did not automatically have this functionality and therefore other software was investigated.

4.3.4.1. Tanagra: percentage contribution to each component

Tanagra v1.4.44 released May 14 2012 (Rakotomalala, 2005) is freely available software which can be used to investigate variance partitioning of the final PLS model (after the sparse procedure had been applied in R). To ensure the model fitting was the same as in R, the final model was run in Tanagra and it agreed that with just one component, 35.75% of the Y variability explained by just 6.4% of the X variability. Tanagra calculates the proportion that each X variable contributes to the first component. Note that the contributions overlap because the variables are correlated. Table 4.4 shows that the first component is made up of 66% of the disease duration, 63% of the symptom duration, 58% of the age at onset of symptoms and 56% of the age at time of diagnosis. The remaining variables contribute very little to this 1st component (<9% each).

Whilst Table 4.4 provides us with an order of variable importance based on the contributions each variable makes to the explainable Larsen score variation, it does not provide us with an estimate of how much Larsen score variation can be explained by the environment and genetics alone because the contributions are not independent. This is further investigated in section 4.3.4.2.

Table 4.4 Tables showing the proportion of variation each variable contributes ('quick' imputation)

>50%	DD*	SymDur#	ageSX $	ageDD^
	0.66	0.63	0.58	0.56

*=Disease duration, #=Symptom duration, $=age at onset of symptoms, ^=age at time of diagnosis

>5% - <50%	rs5029937	rs5029939	rs2230926	rs10499197	rs6932056	rs7752903	rs2071592
	0.085	0.084	0.082	0.076	0.075	0.073	0.055

>2.5% -<5%	rs 2075800	Smoking duration	rs 2844479	rs 5029938	rs 182429	rs 2027276	Smoking Pack year	rs 394581	rs 1800629
	0.043	0.041	0.039	0.032	0.032	0.030	0.027	0.027	0.025

>2%-<2.5%	rs932744	drb1nos2	drb1nos1	rs2242653	rs2009345	rs6933404	rs443198
	0.0228	0.0226	0.0223	0.0216	0.0215	0.0213	0.0212

>1.5%-<2%	rs9366826	rs2327832	rs4535211	rs6927172	rs220704	rs6000570	rs10917214	Alcohol quantity
	0.019	0.0189	0.0187	0.0184	0.0182	0.0164	0.0162	0.0154

>1.2%-<1.5%	rs13242262	rs864745	rs26232	Alcohol use
	0.0145	0.0136	0.0131	0.0127

>1%-<1.2%	rs4777183	rs805292	rs228975	rs11586238	rs12403075	rs10919563	rs2431697
	0.0117	0.0115	0.0111	0.0111	0.0106	0.0104	0.0101

>0.5% - <1%	rs9565072	rs2872507	rs5980742
	0.0088	0.0051	0.005

<0.5%	rs 1076933	rs 7722135	rs 13061519	rs 6473517	rs 8045689	rs 26510	rs 4265819	rs 2256965	rs 2476601
	0.0048	0.0047	0.0044	0.0041	0.0041	0.004	0.0037	0.0029	0.0026

4.3.4.2. MATLAB: Multi-block variance partitioning

Multiblock variance partitioning (Skov et al., 2008) enables quantification of how much model variation is attributable to each type (block) of data. The explainable Larsen score variation can be separated into variation attributable to disease duration and symptom duration (DD&SD), environmental and genetics, in addition to estimating the amount which overlaps between the blocks. One block of variables at a time were fitted in a PLS model and the amount of variation that block alone can explain was calculated. After which the other 'blocks' of variables were added to the model to see what additional variation they can explain. Each block has its own turn of being fitted first in the model and having the other blocks added. The result is that for each block you can estimate the amount of unexplainable variation (not explained by any variables in the model), the amount of unique variation, (the part that block alone explains) and the amount of common variation (the part explained in the block fitted first which is also explained by the blocks later added).

The common and unique variation attributable to the various X blocks of data (DD&SD, environment and genetics) can be partitioned. The method was performed in MATLAB® version 7.13 (R2011b) (The MathWorks Inc., Natick, MA, USA) using the MVP Toolbox (Skov et al., 2008). It also required the installation of the NWAY Toolbox (Andersson and Bro, 2000) available from www.models.life.ku.dk/source/nwaytoolbox.

Fitting DD&SD alone explained 28.1% of the Larsen score variation. However, part of this variation was explained by the other blocks of data, leaving only 18.1% unique to DD&SD. There was much overlap between the DD&SD block and the environment block probably due to age at time of diagnosis and age at onset of symptoms being highly correlated with variables DD&SD. Fitting the genetic block alone explained 16.2% of the Larsen score variation. Almost all of this (15.2%) was unique and not common variation explainable by the other variables (Table 4.5 and Figure 4.5).

Table 4.5 Multi-block variance partitioning ('quick' imputation)

First block fitted	A: Larsen score variation explained by each block when fitted alone (%)	B: Larsen score variation unique to that block i.e. the variation is not common to the other blocks (%). (Percentage unique: B/A*100)
DD&SD	28.1	18.1 (64.5%)
Environment	15.6	10.1 (64.5%)
Genetics	16.2	15.2 (94.1%)

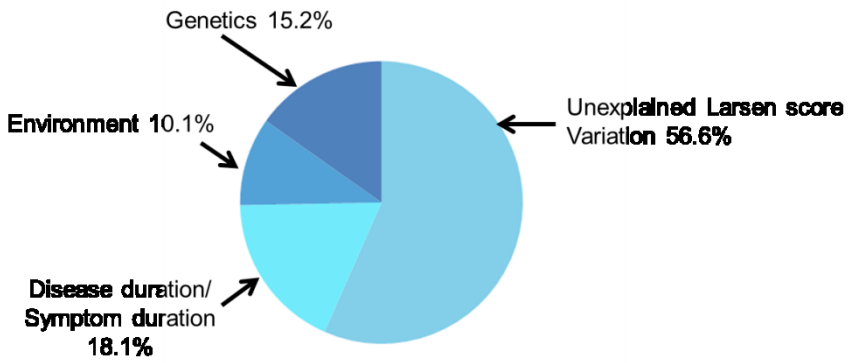


Figure 4.5 Pie chart of Larsen score variation ('quick' imputation)

In summary, of the total Larsen score variation that this model explained (35.75%), excluding variation which was common to other blocks, 15.2% of the Larsen score variability was attributable to genetics alone. Recent estimates of the genetic heritability of radiological damage found possible estimates of 45% (calculated using kinship coefficients) and 58% (calculated using identical by descent) (Knevel et al., 2012a). If SNPs selected due to their known link to auto-immune diseases can explain 15.2%, it is anticipated modelling the 'GWAS SNPs' dataset may be more in the region of the heritability estimates.

It is also worth noting that the sum of the unique parts (18.1, 10.1 and 15.2) did not equate to the total amount of variation (35.75). This was because each block had its own turn at being fitted first in the model and hence its parameter estimates were calculated first, followed by adding in the other blocks to find the unique and common variation. Therefore, the parameter estimates are different depending on the order the blocks are fitted in the model.

4.4. 'Percentage fold' method results when using NIPALS imputation

For comparison with the 'quick imputation' method, the 'all subjects' dataset was imputed using NIPALS imputation and SPLS models were created using the 'percentage fold' method (section 4.2.6). Investigating the optimum number of components for the model resulted in an R^2 of 0.567 with one component which only increased to 0.584 when fitting two components. The R^2-CV for component two was 0.016 suggesting that only one component was required in the model (as it was <0.0975). This was supported by the van der Voet's test which produced a p-value of 0.1460 for the first component. The first component was able to explain 34.9% of the Y variation. Hence all models below were fitted using one component.

4.4.1. Number of variables to extract (NIPALS)

Using the one component model, the maximum R^2-CV was observed at 110 variables (R^2-CV=0.539, Figure 4.6). However it is clear from the plot that this maximum was almost reached at 55 variables (R^2-CV=0.534) and again at 95 variables (R^2-CV=0.535). Given the observable tail off after the 110 variables, it was decided to select 110 variables in each model. It was hoped that if this is too many, they would not be kept in all the folds and therefore not retained in the final model because they had to be selected in 8/10 folds in all 50 runs.

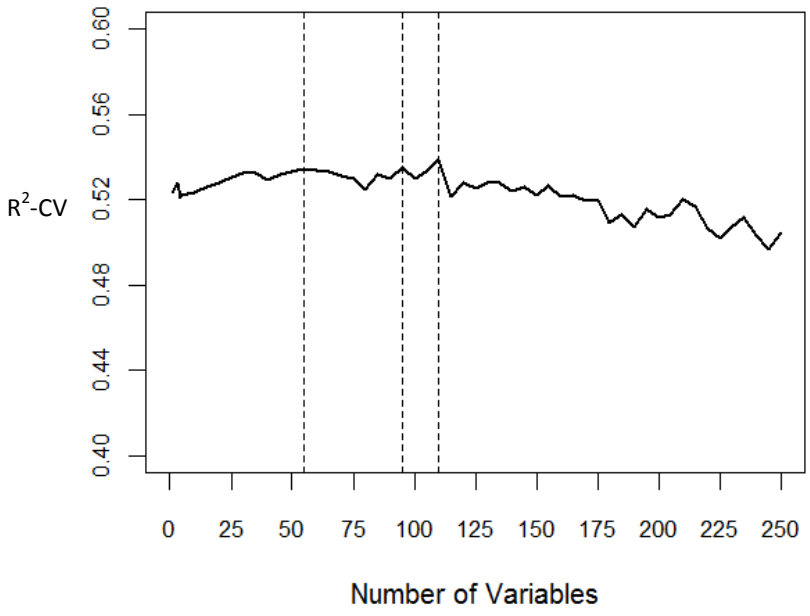


Figure 4.6 Plot of R^2-CV versus number of selected variables (NIPALS)

4.4.2. Selection of the final model (NIPALS)

The model resulted in 59 of the 110 extracted variables being selected in 8/10 folds, in all 50 runs. The rank column in Table 4.6 was calculated as the median rank of each variables loading (from the loading vectors) from the 10 folds within each run followed by taking the median across the 50 runs. The final model is shown along with the sort order (rank), the number of subjects and median Larsen score for each genotype and the ZINB LR p-value for SNPs (section 4.2.5). Only 17 of the 49 SNPs (34.7%) selected were significant at the 5% alpha level. However as shown in section 4.3.2, these p-values are very sensitive to small changes in the data when there are no subjects with a Larsen score=0 and a non-zero genotype.

Table 4.6 Final model using NIPALS imputation

Variable name	Rank	Genotype (0, 1 or 2): N subjects: Median Larsen score			ZINB LR p value	Description
Disease duration	1					Increase in Larsen score for increase in disease duration.
Symptom duration	2					Increase in Larsen score for increase in symptom duration.
Age at onset of	3					Decrease in Larsen score for increase in age at onset of symptoms.
Age at time of diagnosis	4					Decrease in Larsen score for increase in age at disease duration
rs2075800	5	0: 341: 18	1: 424: 33	2: 147: 32	1.31E-07	C/T variant located on chromosome 6 at 31777946. Missense function in HSPA1L.
rs2844479	6.5	0: 375: 20	1: 406: 31.5	2: 131: 38	9.84E-07	G/T variant located on chromosome 6 at 31572956. Unknown function.
rs9366826	8	0: 556: 24.5	1: 320: 31	2: 36: 31	0.0067	C/G variant located on chromosome 6 at 33619184. Intron function in Inositol 1,4,5,-trisphosphate receptor, type 3 (ITPR3).
rs2872507	8.75	0: 298: 21.5	1: 407: 28	2: 207: 33	0.0973	C/T variant located on chromosome 17 at 38040763. Intergenic, 6614 bases downstream of ZPBP2.

Variable name	Rank	Genotype (0, 1 or 2): N subjects: Median Larsen score			ZINB LR p value	Description
rs394581	9	0: 512: 32	1: 345: 20	2: 55: 29	0.1432	C/T variant located on chromosome 6 at 159482771. 5' untranslated region of the TAGAP gene. Associated with the risk of Rheumatoid Arthritis.
rs2071592	11	0: 399: 20	1: 405: 32	2: 108: 33	9.24E-11	A/T variant located on human chromosome 6 at 31515340. Intron variant of NFKBIL1.
Alcohol quantity	11.5					Decrease in Larsen score with increase in alcohol quantity (coded as 0=None, 1=<5 days, 2=5-10 days, 3=> 10 days).
rs26232	12.5	0: 480: 29	1: 342: 27	2: 90: 16	0.0039	C/T variant located on chromosome 5 at 102596720. Intron variant in C5orf30 associated with RA susceptibility and severity.
rs2242653	14	0: 644: 26	1: 240: 32	2: 28: 44.5	4.20E-05	C/T variant location on human chromosome 6 at 31676015. Missense variant in LY6G6F.
rs805292	16.5	0: 648: 25	1: 236: 34	2: 28: 34	1.67E-05	C/T variant located on chromosome 6 at 31690259. 8167 bases upstream of ABHD16A, 4428 bases downstream of LY6G6F, 8167 bases upstream of LY6G6E, 4428 bases downstream LY6G6D, 498 bases upstream of LY6G6D, 1112 bases upstream of C6ORF25, 4808 bases downstream of DDAH2 and 8349 bases downstream of CLIC1.
rs220704	17	0: 703: 26	1: 197: 37	2: 12: 27.5	0.171	A/T variant located on chromosome 6 at 46865758. Intron variant in GPR116.
Alcohol use	17.5					Decrease in Larsen score with drinking alcohol vs. not drinking alcohol.
rs4535211	18	0: 285: 34	1: 435: 26	2: 192: 20	0.0226	A/G variant located on chromosome 3 at 17072997. Intron variant in PLCL2 (Bowes et al., 2012).
rs2431697	18.5	0: 329: 31	1: 422: 27.5	2: 161: 19	0.206	C/T variant located on chromosome 5 at 159879978. Intergenic region, between PTTG1 and microRNA miR-146a associated with Lupus & psoriasis.
drb1nos2	20.5	0: 474: 26	1: 368: 29	2: 70: 39.5	0.939	Increase in Larsen score with presence of 1 or 2 copies of the K-R-A-A shared epitope sequence.
Smoking pack years	22					Decrease in Larsen score for increase in smoking pack years.
rs5029937	23	0: 830: 27	1: 79: 37	2: 3: 31	0.0019	G/T variant located on chromosome 6 at 138195151. Intron variant in TNFAIP3 associated with RA risk (Orozco et al., 2009).
rs182429	23.5	0: 349: 30	1: 447: 26	2: 116: 21.5	0.223	C/T variant located on chromosome 6 at 159469574. 3390 bases upstream of TAGAP (Eyre et al., 2010b).
drb1nos1	29.5	0: 659: 29	1: 235: 26	2: 18: 10	0.0736	Decrease in Larsen score with presence of 1 or 2 copies of A-R-A-A or E-R-A-A shared epitope sequence.
rs7752903	29.5	0: 844: 27	1: 66: 37.5	2: 2: 31	0.0063	G/T variant located on chromosome 6 at 138227364. Intergenic with unknown function. In a 109 kb DNA segment that spans the TNFAIP3 gene possibly related to SLE.
rs932744	30	0: 363: 24	1: 421: 29	2: 128: 35	0.0989	C/G variant located on chromosome 6 at 150390663. 380 bases upstream of ULBP3.
Smoking duration	32.5					Decrease in Larsen score with increase in smoking duration.
rs1076933	34	0: 348: 28.5	1: 416: 28.5	2: 148: 21.5	0.185	A/G variant located on chromosome 22 at 45198494. Intron variant in ARHGAP8 and PRR5-ARHGAP8.
rs2009345	34.25	0: 357: 24	1: 428: 28.5	2: 127: 37	0.123	A/G variant located on chromosome 6 at 150389748. Intron variant in ULBP3.
rs10917214	35	0: 313: 30	1: 437: 27	2: 162: 25.5	0.0809	A/G variant located on chromosome 1 at 22652501. Intergenic with unknown function.
rs5029939	35	0: 828: 27	1: 82: 35.5	2: 2: 31	0.0468	C/G variant located on chromosome 6 at 138195723. Intron variant in the TNFAIP3 associated with RA risk (Orozco et al., 2009).
rs443198	36	0: 396: 24	1: 397: 29	2: 119: 37	2.72E-12	C/T variant located on chromosome 6 at 32190656. Synonymous variant in NOTCH4 associated with activity in RA synovium.
rs228975	38.25	0: 382: 22.5	1: 404: 28	2: 126: 35	0.249	C/T variant located on chromosome 22 at 37542201. Intron variant in IL2RB.
rs2327832	38.5	0: 507: 25	1: 351: 32	2: 54: 24	0.588	A/G variant located on chromosome 6 at 137973068. Intergenic with unknown function.

Variable name	Rank	Genotype (0, 1 or 2): N subjects: Median Larsen score			ZINB LR p value	Description
rs2230926	41.5	0: 827: 27	1: 83: 36	2: 2: 31	0.0444	G/T variant located on chromosome 6 at 138196066. Missense variant in TNFAIP3 associated with RA risk (Orozco et al., 2009).
rs4265819	41.5	0: 753: 26	1: 151: 36	2: 8: 29.5	0.0093	A/G variant located on chromosome 16 at 62617217. Intergenic with unknown function.
rs6000570	41.5	0: 207: 25	1: 458: 26	2: 247: 33	0.647	A/G variant located on chromosome 22 at 37514339. 8736 bases upstream of TMPRSS6 and 7541 bases downstream of IL2RB.
rs13061519	41.75	0: 854: 28	1: 56: 19.5	2: 2: 40	0.272	C/T variant located on chromosome 3 at 173235402. Intron variant in NLGN1.
rs6473517	42	0: 702: 29	1: 197: 21	2: 13: 18	0.082	G/T variant located on chromosome 8 at 84857572. Intergenic with unknown function
rs6933404	42	0: 507: 25	1: 352: 32	2: 53: 24	0.723	C/T variant located on chromosome 6 at 137959235. Intergenic with unknown function
rs8045689	42	0: 488: 26	1: 346: 30	2: 78: 31.5	0.0613	C/T variant located on chromosome 16 at 28988269. Intron variant in SPNS1.
rs12403075	43	0: 722: 29	1: 181: 25	2: 9: 21	0.278	C/T variant located on chromosome 1 at 198721192. Intron variant in PTPRC
rs26510	43	0: 431: 31	1: 387: 23	2: 94: 28	0.343	C/T variant located on chromosome 5 at 96125910. Intron variant in ERAP1
rs7722135	43	0: 580: 25.5	1: 283: 32	2: 49: 36	0.284	C/T variant located on chromosome 5 at 86294669. Intergenic with unknown function.
rs5029938	43.25	0: 852: 28	1: 60: 18.5	2: NA: NA	0.594	C/T variant located on chromosome 6 at 138195633. Intron variant in TNFAIP3 associated with RA risk (Orozco et al., 2009)
rs2476601	43.5	0: 646: 26	1: 234: 32.5	2: 32: 44	0.117	A/G variant located on chromosome 1 at 114377568. Missense variant in PTPN22 associated with RA risk and possibly ACPA positivity.
rs4777183	45.75	0: 282: 30.5	1: 434: 27.5	2: 196: 20	0.428	C/T variant located on chromosome 15 at 70004775. Intergenic with unknown function.
rs10499197	48	0: 846: 27	1: 64: 36.5	2: 2: 31	0.072	C/T variant located on chromosome 6 at 138132516. Intergenic with unknown function.
rs6932056	48	0: 837: 27	1: 73: 36	2: 2: 31	0.035	C/T variant located on chromosome 6 at 138242437. Intergenic with unknown function.
rs10919563	48.75	0: 729: 29	1: 174: 25.5	2: 9: 21	0.173	A/G variant located on chromosome 1 at 198700442. Intron variant in PTPRC
rs6927172	52	0: 524: 26	1: 337: 32	2: 51: 24	0.555	C/G variant located on chromosome 6 at 138002175. Intergenic with unknown function.
rs2027276	52.5	0: 852: 28	1: 60: 19.5	2: NA: NA	0.657	G/T variant located on chromosome 6 at 138174328. Intron variant in LOC100130476.
rs2256965	53.5	0: 447: 24	1: 368: 30	2: 97: 33	5.26E-09	C/T variant located on chromosome 6 at 31555130. Intron variant of LST1: Leukocyte-specific transcript 1 protein
rs864745	54	0: 291: 30	1: 421: 27	2: 200: 26.5	0.233	A/G variant located on chromosome 7 at 28180556. Intron variant in JAZF1 which has previously been associated with Type II diabetes.
rs2002842	54.25	0: 328: 25	1: 416: 27	2: 168: 32	0.084	A/C variant located on chromosome 18 at 76409597. Intergenic with unknown function.
rs11586238	55	0: 539: 30	1: 318: 21	2: 55: 26	0.188	C/G variant located on chromosome 1 at 117263138. Intergenic with unknown function however possibly associated with RA risk (Raychaudhuri et al., 2009)
rs12137270	62.25	0: 539: 30	1: 318: 21.5	2: 55: 26	0.205	C/T variant located on chromosome 1 at 117264336. Intergenic with unknown function.
rs7234029	63	0: 640: 29	1: 239: 24	2: 33: 18	0.033	A/G variant located on chromosome 18 at 12877060. Intron variant in PTPN2.
rs3788013	63.5	0: 315: 32	1: 432: 26	2: 165: 26	0.157	A/C variant located on chromosome 21 at 43841328. Intron variant in UBASH3A
rs17810546	64.5	0: 723: 29	1: 170: 25	2: 19: 18	0.116	A/G variant located on chromosome 3 at 159665050. Intron variant in AK097161 possibly associated with celiac/juvenile idiopathic arthritis

Note: For tabulation purposes, results imputed to be a number containing a decimal between 0 and 2 using NIPALS imputation were rounded to the nearest genotype and counted in that category. However, the data were analysed as the imputed decimal value in the model and calculation of the p-value.

4.4.3. Assessing predictive ability of the model (NIPALS)

The final model explained 34.9% of the Y variability with just 6.24% of the X variability. A correlation of 0.592 was achieved between the predicted and actual Larsen score after refitting the model on the same patients used to form the model (Figure 4.7).

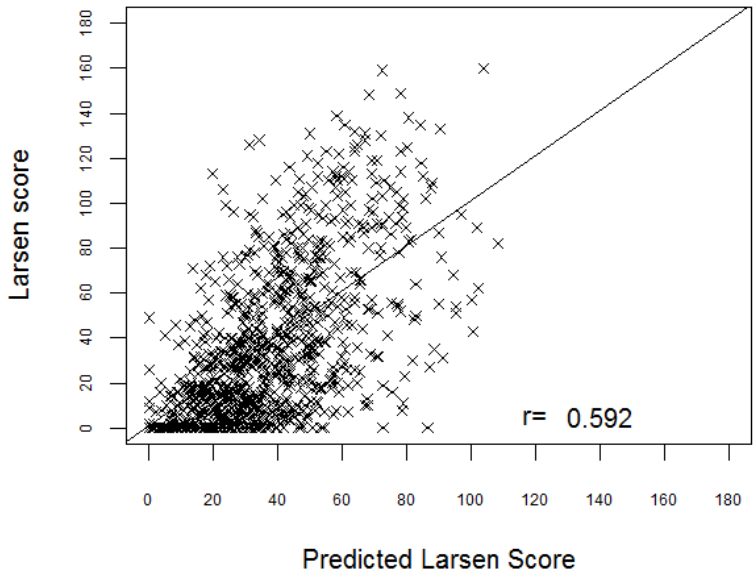


Figure 4.7 Scatter plot of actual vs. predicted Larsen score (NIPALS)

The mean absolute difference between the actual Larsen score and the predicted one was 21.93 (Median=18.38, STD=17.62, Minimum=0 and Maximum=94.88).

4.4.4. Variance partitioning (NIPALS)

Using the multi-block variance partitioning method (Skov et al., 2008) described in section 4.3.4.2, the NIPALS model explained slightly less Larsen score variation than the 'quick' imputation model (34.9% versus 35.75%). Table 4.7 revealed that almost the same percentage of unique Larsen score variation was attributable to DD&SD (18.3% versus 18.1%) and there was a similar overlap between the DD&SD and environmental blocks. Fitting the genetic block alone explained only 13.9% of the Larsen score variation (compared to 16.2% in the 'quick' imputation model), but again nearly all of it (13.2%) was unique and not common to the other variables.

Table 4.7 Multi-block variance partitioning (NIPALS)

First block fitted	A: Larsen score variation explained by each block when fitted alone (%)	B: Larsen score variation unique to that block i.e. the variation is not common to the other blocks (%). (Percentage unique: B/A*100).
DD&SD	28.1	18.3 (65.2%)
Environment	16.3	10.5 (64.5%)
Genetics	13.9	13.2 (94.7%)

4.5. Summary

A SPLS multivariate modelling strategy was investigated on the 'all subjects' dataset (N=912). 368 SNPs and 19 environmental variables were reduced down to the most predictive of the Larsen score. Two separate imputation methods were investigated. 'Quick' imputation assigned 0's for all missing SNPs, the mean for all missing continuous variables and the mode for all missing environmental categorical variables. NIPALS imputation assigned values iteratively using lines of best fit through multiple bivariate regression models. Using the SPLS model creation process 'percentage fold' method described in section 4.2.6, only one component was required in all models investigated. The number of variables to extract in each run and fold of the model was determined according to the number which explained the largest proportion of the Larsen score variation under CV. The number of variables which were extracted in 8 out of the 10 CV folds in all 50 runs were retained for the final model.

Exploration into how SPLS copes with missing data discovered a bias towards selection of SNPS with lots of missing data when imputation using the NIPALS algorithm was used. This had been identified by other authors and was reported to be possibly due to NIPALS artificially reducing the variation in the data (Pedreschi et al., 2008). Hence, this would be a problem when using any modelling methods. To avoid excessive variance reduction, it was recommended to exclude variables with greater than 20% data missing from the modelling (Trygg and Wold, 2002). When SNPs with>20% missing data were excluded the average missing data was comparable across the SNPs selected and not selected. Therefore for all subsequent models, SNPs with >20% missing data will be removed prior to modelling.

Modelling the Larsen score divided by the disease duration and log (Larsen score divided by disease duration), with or without excluding the zero Larsen scores was explored. The predictive ability was not as good as modelling the Larsen score untransformed. It is hypothesised that the reduction in predictive ability was because the effect of disease duration is likely not to be linear over time. Subjects in their first three years after diagnosis can progress much quicker than would be expected for a three year period from 10 to 13 years or 40 to 43 years disease duration. This may introduce a non-linear relation and hence additional variability into the modelling process which was not predictive of Larsen score severity. Therefore, it was decided to model Larsen score without any transformation.

Similar models were obtained using the 'quick' & NIPALS imputation methods. For the 'quick' imputation method, 95 variables were extracted in each fold and run which resulted in 58 variables selected in at least 8 out of 10 folds in all 50 runs. These 58 variables were retained for the final model. For NIPALS imputation, 110 variables were extracted which resulted in 59 selected for the final model. 54 of these variables were the same in both models. It is also reassuring to note that the choice of number of variables to extract did not appear to have an impact on the two models. The 'quick' imputation model had four variables (rs13242262, rs1800629, rs5980742 and rs9565072) not observed in the NIPALS model however they are ranked 35th or lower in the selection of variables. The NIPALS model had five variables (rs12137270, rs17810546, rs2002842, rs3788013 and rs7234029) not selected in the 'quick' imputation model. Reassuringly, these are five of the six last SNPs selected. Hence the models were very similar.

The correlation between the actual Larsen score and the predicted Larsen score was; r=0.599 ('quick') and r=0.592 (NIPALS). The SPLS model was able to explain; 35.74% ('quick') and 34.9% (NIPALS) of the Y variation. Using multi-block variance partitioning; 15.2% ('quick') and 13.2% (NIPALS) of the unique Larsen score variation was explained by genetics.

It is well documented that using the mean to impute values can add false precision and reduce the estimated true variation of the sample (Schafer and Graham, 2002). It is hypothesised that using a zero for all missing SNPs per the 'quick' imputation method may have the same effect. Using multi-block variance partitioning, the 'quick' imputation method attributed 15.2% of the variation to genetics compared to the smaller 13.2% for the NIPALS imputation modelling. If the 'quick' imputation method underestimates the total variation because all missing data is assigned a zero, then the percentage that the genetics explains could be an over estimation. Prior to modelling the 'GWAS SNPs' dataset, the method of imputation will be explored further to incorporate imputing SNPs in LD (Chapter 5).

5. Investigation into imputation methods – 'GWAS SNPs' dataset

5.1. Aims

The aim of this chapter is to:

- Evaluate SPLS model fitting using various imputation methods applied to a small section of the 'GWAS SNPs' dataset.
- Investigate the reproducibility of SPLS model fitting and variable selection.
- Further explore the SPLS methodology in terms of the number of variables to extract and number of runs and folds a variable has to be selected in to be kept in the final model.

5.2. Methods of model fitting, region selection and imputation

Prior to modelling the 'GWAS SNPs' dataset, the method of imputation of any missing SNP data will be investigated to ensure SNP selection is due to their predictive ability of the Larsen score and the model is not sensitive to the imputation method used. All SNPs require imputation as they are being fitted in a single multivariate model (section 3.8). This research does not investigate the imputation of non-genotyped SNPs as primary focus is to explore whether SPLS could select SNPs correlated with severity rather than investigate causal SNPs in a region.

Four imputation methods are explored. Two methods ('quick' imputation and NIPALS) are previously described in section 4.2.2. Two additional methods (IMPUTE2 and PLINK) are described in sections 5.2.3 and 5.2.4 respectively.

5.2.1. 'Percentage fold' method for subset of 'GWAS SNPs' dataset

The general method of model fitting is the same as described in section 4.2.6 with the exception of the number of folds used and the number of folds a variable has to be selected in to be retained for the final model. Only 380 subjects were included in this chapter due to patient identification discrepancies for 14 subjects which were resolved by chapter 6. Five-fold CV (instead of 10-fold) will be used, resulting in the training models consisting of $4/5^{ths}$ of the data (304 subjects) and each model being tested on the remaining 76 subjects.

The SPLS analysis requires each dataset to have any SNPs with >20% missing data removed (section 4.2.3). In addition, in the CV process of the SPLS models, the model fails if there is no variation in the training sample for any single SNP. The process could therefore fail on one run of the CV models, if it so happened that 4 of the 5 folds (304 subjects) were all 0 (most frequent homozygous genotype) or very close to 0 (in the case of NIPALS imputation using bivariate regressions which are not constrained to whole numbers). It is therefore necessary to amend the mixOmics macros to exclude individual SNPs from certain folds, if they have no variation in the training set to prevent the process failing. After exploring various alternatives the following strategy will be used:

1. On the training set (4/5 folds=304 subjects), count the number of subjects with results between 0 and 0.5. 0 is a subject with most frequent homozygous genotype, however, if using NIPALS imputation derived using bivariate regressions it is not constrained to be 0, 1 or 2. Hence the programming allows for imputations close to 0 but not exactly 0. Less than 0.5 is used as this is where NIPALS is suggesting the SNP is more likely to be 0 than to be 1.
2. Remove any SNP from that training set if >92% of the 304 subjects (280 subjects) all have results between 0 and 0.5. Model creation when less than 24 subjects have a genotype of 1 or 2 is likely to be unstable when applied to other populations. Whilst this may remove some rare SNPs from the model fitting, it is done on a fold by fold basis and so they will still be included in other folds if they have more subjects with minor alleles.

Other criteria for exclusion were examined, however the above condition rarely fails whilst still allowing SNPs with very few rare genotypes to be included. The above restriction on the SNPs available for modelling, results in slightly different SNPs being available for selection in each imputation dataset.

These changes which were made to the mixOmics macro can be found in Appendix D.

5.2.2. Region selection

The HLA-DRB1 region on chromosome 6 has previously been identified as an area of key interest for Rheumatoid Arthritis severity and susceptibility (Tezenas du Montcel et al., 2005, Michou et al., 2006). For this reason and that it is in high LD, it was selected as an appropriate area to use from the entire genome to test the methods of imputation. The GWAS study was built in Hg18 - build 36 and therefore all positions are referencing that structure. Using the region on chromosome 6 from 30-35 million base pairs, all SNPs between rs9259806 and rs1053049 (nearest SNP above 35000000) were selected. The 3504 SNPs in this region were exported into Haploview version 4.2 (Barrett et al., 2005) to view the LD structure of the GoRA 'GWAS SNPs' dataset collected in this region (Figure 5.1).

Reviewing sections of the Figure 5.1 in zoomed in detail, two areas close together were selected as regions of lower and higher LD. The higher LD section consisted of 374 SNPs from rs9267956 (32321616 bp) to rs7745656 (32788948 bp). The lower LD region consisted of 471 SNPs from rs12206131 (31521989 bp) to rs11966200 (31945045 bp). The higher LD region also included the HLA-DRB1 region 32655000 to 32665000 bp. The two-by-two correlations between SNPs based on genotype allele counts were exported from PLINK (Purcell et al., 2007) as an approximate estimate of the strength of LD. The absolute correlation in the high LD dataset was much larger (mean=0.44, median=0.29) than that observed in the lower LD region (mean=0.06, median=0.03). Correlations were used instead of traditional r^2 and D' statistics to measure LD because they were faster to calculate.

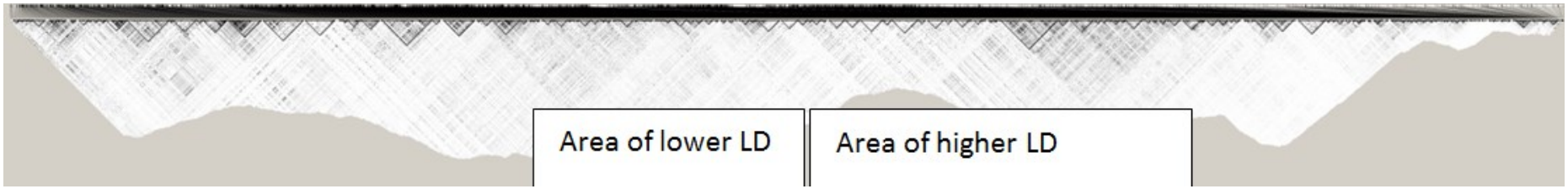


Figure 5.1 LD pattern for the GoRA 'GWAS SNPs' dataset from 20-25 M BP.

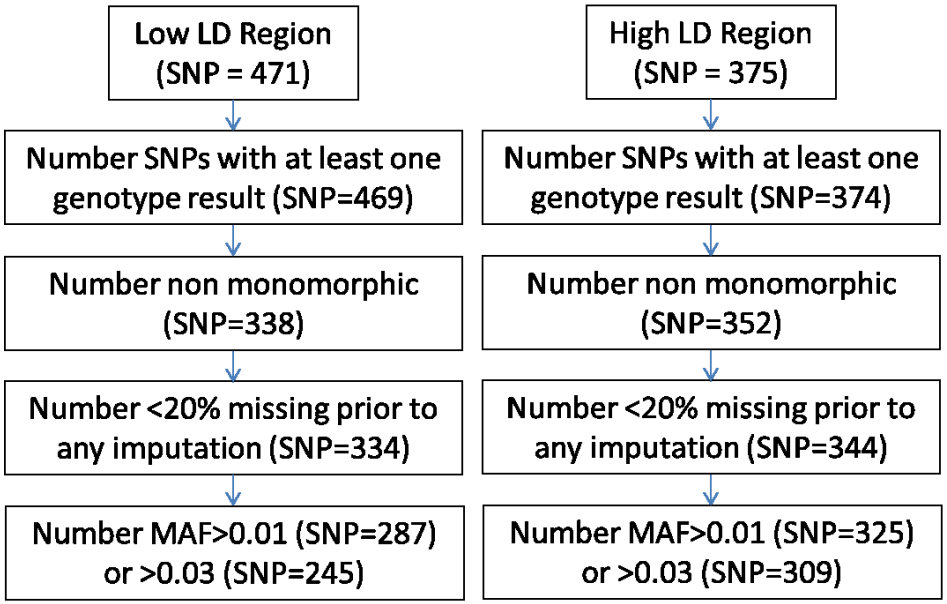


Figure 5.2 Flow diagram of exclusion of SNPs by reason

Using the methods described in 5.2.1, SNPs with zero variation (monomorphic), SNPs with >20% missing genotypes and SNPs with insufficient minor allele frequency (MAF) based on that fold and run of the data were excluded from the analysis. Figure 5.2 shows a high level summary of the regions analysed prior to imputation. As different imputation methods are applied prior to analysis, the number of SNPs with sufficient data to be analysed varies by imputation method. Table 5.1 shows the exact number of SNPs used by imputation method.

The SNP data were merged with the 19 environmental variables described in 4.2.2. SNPs in these two regions are imputed and analysed using SPLS to enable comparison of the sensitivity of the PLS models to various imputation methods.

5.2.3. IMPUTE2 version 2.2.2

IMPUTE2 http://mathgen.stats.ox.ac.uk/impute/impute_v2.html (Howie et al., 2009) uses observed haplotypes and compares them against a reference set of haplotypes. For this analysis, the reference data was taken from the 1000 genomes project (sequence data from March 2010 and phased haplotypes from June 2010) and the HapMap 3 project (release number 2, February 2009) using NCBI genome build 36 (hg18) co-ordinates. The low LD region (3.15e7 to 3.20e7) and the high LD region (3.23e7 to 3.28e7) were imputed using the options as recommended on the IMPUTE2 website as follows:

- Ne: Effective size of the population: 20000 was recommended for use
- Call_thresh 0.5: Reduced the default threshold from 0.9 to 0.5. This means that SNPs with missing data were only assigned a predicted value for the three possible genotypes (0, 1 or 2) when one of the predictions was more than 50% likely.
- Pgs_miss: Only replace the missing genotypes at typed SNPs. Hence any present genotypes were reprinted in the output file.
- -os 2: Only output SNPs which were measured in the GoRA 'GWAS SNPs' dataset (not SNPs from the reference dataset in the same area).

The region of interest was exported from PLINK as a .PED and .MAP file, converted into .GEN file using the software Gtool (http://www.well.ox.ac.uk/~cfreeman/ software/gwas/gtool.html) and read into IMPUTE2. IMPUTE2 was used to provide a dataset retaining the input (GoRA GWAS) SNPs only. Any non-missing data was carried through with the correct genotype receiving a probability of 1 and the other genotypes a probability of 0. For any missing SNP data, IMPUTE2 assigned a probability of each genotype being the true genotype. SPLS modelling cannot be performed using a dataset where each subject has multiple possible responses for each SNP and each of these responses has a probability of being the true response. Instead, SPLS requires a single value for each SNP, a best estimate of what the missing result would be.

It was therefore decided to convert the data back through Gtool (.GEN to .PED and .MAP) assigning each genotype as the one IMPUTE2 predicted with a probability greater than 0.5. Any SNPs with no genotype predicted more than 50% likely to be the correct result were left as missing.

A second option could have been to calculate a score between 0 and 2 which consisted of the weighted genotype probabilities. However, for some SNPs IMPUTE2 assigned each genotype equal probability. For example, the SNP had a 33% chance of being a 0, 33% chance of being a 1 and a 33% chance of being a 2 (i.e. each genotype is equally likely). Using weighted probabilities would assign a 1 (0.33*0 + 0.33*1+0.33*2 = 1). This appears to be quite misleading as all the genotypes are all equally likely to be correct, however heterozygous is assigned. Instead only using IMPUTE2 to assign genotypes with a probability greater than 0.5 will utilise the external reference panel approach, but will leave any SNPs missing if they do not have enough information to produce a good prediction using this method.

Once a predicted genotype is selected in Gtool, files are then converted through PLINK (from .PED and .MAP to .BID, .BAM and .FAM and then to .RAW) to be read into SAS® for merging on of the environmental data and comparison with the non-imputed dataset.

This comparison revealed that 300 SNPs remained for the high LD region. This was comparable to the 315 SNPs retained in the non-imputed high LD dataset after the SNPs with insufficient variation had been removed. 164 (52.6%) had at least some imputation performed using this method. Four SNPs (rs35603463, rs9271622, rs9268181 and rs9274791) had more than 100 patient's genotypes imputed (198, 182, 138 and 137 respectively). Excluding these four, the remaining 160 SNPs had a mean = 9.35, median =3, STD =14.8, min=1 and max=91 genotypes imputed.

Comparison of the low LD region pre and post imputation revealed that 241 SNPs remained after imputation in IMPUTE. This was comparable to the 248 SNPs retained in the non-imputed low LD dataset after those with insufficient variation were removed. As is to be expected in a region with lower LD, fewer SNPs, 85 (32.1%), were able to have some imputation performed. Two SNPs (rs769178 and rs2523675) had 139 and 118 patient's genotypes imputed respectively. After the exclusion of these two, the remaining SNPs had a mean =7.8, median=3, STD=13.4, min=1 and maximum = 89 genotypes imputed.

Using this method in IMPUTE2, it was unable to impute all missing values in the data. To enable SPLS modelling, post using IMPUTE2 the remaining missing genotypes were imputed using a NIPALS algorithm per the standard approach of PLS. Although this may make the dataset similar to the NIPALS imputation, it is hoped that the prior imputation using the reference panel would mean the datasets are different enough to compare if the IMPUTE2 step is beneficial to the PLS modelling.

5.2.4. PLINK version 1.07

PLINK (http://pngu.mgh.harvard.edu/purcell/plink/) (Purcell et al., 2007) is a non-computationally intense method which imputes SNPs using the concept of multi-marker tagging and is designed only for use in an exploratory manner. However, because it is very quick to impute, it was intended to provide a fourth method for exploration of how sensitive the chosen SNPs are to the imputation method used. The reference dataset used was the Phase 2 HapMap: CEU founders (release 22), containing data from 60 individuals and 2.3 million SNPs in build 36 (Hg18). Any SNPs not on the correct strand were flipped and there were no SNPs which were in the selected region location under a different base pair position. For the high LD region, 75 (24.0%) SNPs had at least some imputation performed using this method. Three SNPs (rs9268181, rs28594633 and rs28877027) had more than 100 patient's genotypes imputed (138, 201 and 317 respectively). Excluding these three, the remaining 72 SNPs had a mean = 9.5, median =3, STD =14.9, min=1 and max=66 genotypes imputed. It was a similar picture for the low LD region with 68 (25.7%) SNPs having at least some imputation performed. Two SNPs (rs769178 and rs9267481) had 105 and 131 patient's genotypes imputed respectively. After the exclusion of these two, the remaining 66 SNPs had a mean =7.0, median=3, STD=10.6, min=1 and maximum = 53 genotypes imputed.

This quicker method of imputation does not impute as much data as the IMPUTE2 software and is documented for exploratory use only. Given there were only 60 individuals in the available reference panel for PLINK, compared to greater than 1000 in the IMPUTE2 software, IMPUTE2 appears the better method. It was therefore decided to only test the 'quick', NIPALS and IMPUTE2 imputation.

5.3. 'Percentage fold' method results on subset of 'GWAS SNPs' dataset

In the following work six datasets were available for exploration:

- High LD quick: Region of high LD imputed using the 'quick' imputation
- Low LD quick: Region of low LD imputed using the 'quick' imputation
- High LD NIPALS: Region of high LD imputed using the NIPALS imputation
- Low LD NIPALS: Region of low LD imputed using the NIPALS imputation
- High LD IMPUTE: Region of high LD imputed using IMPUTE2 software followed by NIPALS
- Low LD IMPUTE: Region of low LD imputed using IMPUTE2 software followed by NIPALS

5.3.1. Reproducibility of PLS

As the 'GWAS SNPs' dataset used in this chapter was more highly correlated than the 'all subjects' dataset used in chapter 4, it was of interest to explore whether the SPLS model is truly reproducible even in the presence of correlated variables. Using the high LD 'quick' imputation model as an example, the 380 subjects and 334 X variables were run through a SPLS model selecting 45 variables for each fold and only retaining a variable if it was selected in 2/5 folds in all 50 models. The variable selection criteria (45 variables and 2/5 folds) was lower than previously used in chapter 4 but chosen arbitrarily before further investigation shown in section 5.3.2. The same model fitting criteria was repeated three times with a full 50 runs done each time. It was hoped that running the exact same model over again, would result in the same variables being selected because in each case, the ordering of the variables was averaged over 50 runs of 5-fold CV. However, the different random allocation of subjects to folds could lead to some variation in the variables chosen.

The three repetitions selected 33, 33 and 32 variables for the final model. 31 variables were consistent in all three models. Two variables were selected in two models, and one variable was only selected in one model (but it was the last variable selected). Investigation revealed the reason for the inconsistency was that the variables were not being selected in 50/50 runs. The random selection of patients into folds, in that particular run, caused the variable to not be extracted in at least 2/5 folds, however, it had been selected in 2/5 folds in the 49 other runs. This could indicate that using n/N folds and insisting on the selection in all runs, may not be the best way of choosing variables for the final model and an alternative approach is examined in section 6.4.1. With this number of variables, the models were quite similar hence the methodology was not amended at this stage, however, this was reassessed when modelling the entire 'GWAS SNPs' dataset.

5.3.2. Investigating the variable selection criteria

The number of components and the number of variables to extract in each fold of the data were determined using methods described in 4.2.6 for each of the imputed datasets. Plots of the R^2-CV for all models selecting between one and 150 variables revealed R^2-CV peaks in the regions described in Table 5.1. Often there was no clear peak (with little gain or loss in R^2-CV between four and 25 variables) hence a range was provided to enable selection of one value to take forward for all models. All models suggested the use of one component. The variation in the number of SNPs available for modelling shown in Table 5.1 was due to the different methods of imputation (see 5.2.1).

Table 5.1 Summary of imputation method parameters

	High LD			Low LD		
	'Quick'	NIPALS	IMPUTE	'Quick'	NIPALS	IMPUTE
N subjects	380	380	380	380	380	380
Number environmental variables	19	19	19	19	19	19
Number of SNPs	315	317	300	248	249	241
Number of components	1	1	1	1	1	1
Number of variables to select (peak in R^2)	4-25	4-35	4-25	4-30	4-30	4-25

Previously when using 10 folds, a variable had to be extracted in 8/10 folds in each run to be retained in the final model. Now only five folds were being used, it was felt 4/5 folds would be too strict on rarer SNPs, which may not be able to be included in all folds due to sparse variation in the data and the smaller sample size being used. The high LD 'quick' dataset was used to explore changes in model selection when extracting different numbers of variables and only keeping them in the final model if they appear in 40% (2/5) and 60% (3/5) of the folds.

Table 5.2 reveals that it is a trade-off between the percentage of folds a variable is required to be in, versus the number of variables extracted each time from the fold. For example, extracting 30 variables using the 2/5 fold criteria (Model 1), results in an almost identical model to extracting 40 variables and using the 3/5 fold criteria (Model 4). Extracting 40 variables and 2/5 fold criteria (Model 3) is similar to extracting 45 variables using 3/5 fold criteria (Model 6). If 50 variables and 3/5 folds had been used, this would be somewhere between Model 6 and Model 7 and would be similar to Model 3.

Of particular interest, was the lack of increase in R^2-CV when fitting more variables, perhaps identifying that none of the SNPs explored have a particularly large impact on the Larsen score prediction. To investigate further, two sets of models (firstly extracting 45 variables and secondly extracting 20 variables), were produced using the 2/5 fold criteria using all three imputation methods for the low and high LD regions. The 20 variable models performed better in terms of prediction, compared to the 45 variable models, for all imputation methods and for SNPs in low and high LD (data not shown). This indicated that whilst too few variables may miss important predictors, too many variables can lead to over fitted models which perform poorly in CV.

In order to explore if different imputation methods performed better or worse than each other, it was decided to extract 25 variables and only keep them in the final model if they appear in 2/5 folds in all 50 runs. This intended to keep the important predictive variables, whilst not making the model over fitted and provide a comparison of all of the different imputation methods. The variable selection methods were revisited in section 6.4.1. Models comparing the different methods of imputation, based on these criteria, are shown in section 5.3.3.

Table 5.2 Investigation using high LD 'quick' imputation data into variable selection conditions

	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6	Model 7
N extracted	30	30	40	40	45	45	60
N folds required	2/5	3/5	2/5	3/5	2/5	3/5	2/5
The order that the variables were chosen in has been amended so they can be compared. The top 17 variables were the same in all models. In addition, with the exception of the most restrictive model (2), the top 21 variables are the same across all models.	DD	DD	DD	DD	DD	DD	DD
	SymDur	SymDur	SymDur	SymDur	SymDur	SymDur	SymDur
	Age Diag	Age Diag	Age Diag	Age Diag	Age Diag	Age Diag	Age Diag
	Age onset	Age onset	Age onset	Age onset	Age onset	Age onset	Age onset
	smkdur	smkdur	smkdur	smkdur	smkdur	smkdur	smkdur
	packyear	packyear	packyear	packyear	packyear	packyear	packyear
	rs3117133	rs3117133	rs3117133	rs3117133	rs3117133	rs3117133	rs3117133
	rs2894249	rs2894249	rs2894249	rs2894249	rs2894249	rs2894249	rs2894249
	rs3129941	rs3129941	rs3129941	rs3129941	rs3129941	rs3129941	rs3129941
	rs6904320	rs6904320	rs6904320	rs6904320	rs6904320	rs6904320	rs6904320
	rs3129907	rs3129907	rs3129907	rs3129907	rs3129907	rs3129907	rs3129907
	rs3129871	rs3129871	rs3129871	rs3129871	rs3129871	rs3129871	rs3129871
	rs2395173	rs2395173	rs2395173	rs2395173	rs2395173	rs2395173	rs2395173
	rs3135338	rs3135338	rs3135338	rs3135338	rs3135338	rs3135338	rs3135338
	Alcohol Q	Alcohol Q	Alcohol Q	Alcohol Q	Alcohol Q	Alcohol Q	Alcohol Q
	rs9271348	rs9271348	rs9271348	rs9271348	rs9271348	rs9271348	rs9271348
	Alcohol use	Alcohol use	Alcohol use	Alcohol use	Alcohol use	Alcohol use	Alcohol use
	rs2395182		rs2395182	rs2395182	rs2395182	rs2395182	rs2395182
	rs9501626		rs9501626	rs9501626	rs9501626	rs9501626	rs9501626
	rs2856705		rs2856705	rs2856705	rs2856705	rs2856705	rs2856705
	rs9275245		rs9275245	rs9275245	rs9275245	rs9275245	rs9275245
			rs7775228		rs7775228		rs7775228
	smoke		smoke		smoke	smoke	smoke
			rs3117116		rs3117116	rs3117116	rs3117116
			rs9268104		rs9268104		rs9268104
			rs2395175		rs2395175		rs2395175
			rs28891406		rs28891406		rs28891406
			rs9268118		rs9268118		rs9268118
			rs6457617		rs6457617		rs6457617
			rs6910071		rs6910071		rs6910071
					rs9267992		rs9267992
					rs3104369		rs3104369
					rs13204672		rs13204672
							rs6903608
							rs9268880
							rs3763309
							rs3793127
							rs9268585
							rs17212420
							rs9368726
							rs2395163
							rs5000634
							rs3129922
							rs9273448

DD=disease duration,
SymDur=symptom duration
Age onset=age at disease onset
Smkdur= Smoking duration
Packyear= Smoking pack years
Alcohol Q= alcohol quantity
Smoke= Smoking status

5.3.3. Comparison of imputation methods

The six imputation datasets were processed using SPLS models using 5-fold CV and extracting the top 25 variables. Models were run 50 times and any variables appearing in 2/5 folds in all 50 runs were selected for the final model. These final selected models were then re-run on the entire dataset to obtain summaries of the prediction accuracy.

The imputation models performed similarly for both the high and low LD regions. The NIPALS and 'quick' imputation model performed slightly better than the IMPUTE2 model in the high LD dataset and the IMPUTE2 model performed slightly better than the 'quick' imputation and NIPALS in the low LD dataset (Table 5.3).

In conclusion, the method of variation selection appeared more important than the imputation method used. This could be because any SNPs with >20% missing are removed from the analysis.

Table 5.3 Comparison of predictive accuracy of the different imputation method models

	High LD			Low LD		
	'Quick'	NIPALS	IMPUTE2	'Quick'	NIPALS	IMPUTE2
Number of X variables	19	18	20	23	23	19
Correlation (Act versus Pred)	0.558	0.563	0.554	0.570	0.570	0.588
Mean abs (Act-Pred)	24.83	24.78	24.91	24.51	24.51	24.24
Median abs (Act-Pred)	22.14	22.32	22.21	21.35	21.36	21.00
Min abs (Act-Pred)	0.004	0.005	0.027	0	0	0
Max abs (Act-Pred)	98.14	97.61	102.39	85.79	85.82	84.55

Note: Number of X variables is the number selected for the final model. Act =Actual Larsen score, Pred=Predicted Larsen score, abs=absolute value.

When examining the variables selected in each of the imputation methods for the high LD region, (Table 5.4) the variables were almost identical across the three models. Just rs2395182, Smoking and rs9501626 are not consistently selected across the three models. In models where these were included they were the last three variables selected and were not chosen in all of the folds (demonstrated by their median order being >25).

Table 5.4 Comparison of variables selected for imputation methods using the High LD region

	'Quick': β (order[a])	NIPALS: β (order[a])	IMPUTE: β (order[a])
Intercept	67.46 (0)	66.87 (0)	67.34 (0)
Disease duration	0.399 (1)	0.409 (1)	0.393 (1)
Symptom duration	0.352 (2)	0.361 (2)	0.347 (2)
Age at time of diagnosis	-0.26 (3)	-0.266 (3)	-0.256 (3)
Age at onset of symptoms	-0.255 (4)	-0.261 (4)	-0.251 (4)
Smoke duration	-0.08 (5)	-0.082 (5)	-0.078 (5)
Smoking pack years	-0.072 (6)	-0.074 (6)	-0.071 (6)
rs3117133	-2.741 (6.5)	-2.735 (7)	-2.075 (19)
Alcohol quantity	-1.181 (14)	-1.211 (13)	-1.163 (14)
rs2894249	-2.446 (14)	-2.508 (15)	-2.408 (13)
rs3129941	-2.446 (14)	-2.508 (15)	-2.408 (13)
rs6904320	-2.446 (14)	-2.508 (15)	-2.408 (13)
rs3129907	-2.408 (15)	-2.573 (14)	-2.463 (7)
rs2395173	-1.89 (16.5)	-1.938 (16.5)	-1.86 (15.5)
rs3135338	-1.89 (16.5)	-1.938 (16.5)	-1.86 (15.5)
rs3129871	-1.857 (17)	-1.931 (15)	-1.828 (15)
rs9271348	-2.198 (17)	-2.253 (17)	-2.163 (18)
Alcohol use	-2.395 (20)	-2.455 (25)	-2.357 (19)
rs2395182	-1.914 (171)		-1.884 (22)
Smoking status	-1.257 (171)		-1.238 (165.5)
rs9501626		-2.761 (171)	-2.65 (166)

Note: [a] order consists of ranking the variables according to the amount of Larsen score variation explained, if outside the top 25 for that fold in that run, then they are given an order equal to the total number of variables (equal last as not selected in the model). The median rank (order) is then calculated over the folds within a run and then over the runs. Hence, a median order >25 implies for some folds these variables were not selected.
Although parameter estimates (beta coefficients) are presented here for comparison across models they require careful interpretation as the size of them will depend on how correlated that variable is with other variables in the model.

Similarly for the low LD region, the variables selected in the 'quick' imputation and NIPALS model were identical (Table 5.5). The IMPUTE2 model consisted of the same variables in the other two with the exception of rs1266076, rs805301, rs813115 and rs805302 which were four of the last five variables selected.

Table 5.5 Comparison of variables selected for imputation methods using the low LD region

	'Quick': β (order[a])	NIPALS: β (order[a])	IMPUTE: β (order[a])
Intercept	58.58 (0)	58.51 (0)	64.6 (0)
Disease duration	0.373 (1)	0.373 (1)	0.412 (1)
Symptom duration	0.329 (2)	0.329 (2)	0.364 (2)
Age at time of diagnosis	-0.243 (3)	-0.242 (3)	-0.269 (3)
Age at onset of symptoms	-0.238 (4)	-0.238 (4)	-0.263 (4)
Smoking duration	-0.074 (5)	-0.074 (5)	-0.082 (5)
rs3130617	-2.317 (5.5)	-2.315 (5.5)	-2.564 (5)
rs2857597	-2.304 (6)	-2.303 (7)	-2.55 (6)
rs28366162	-4.883 (7)	-4.88 (7)	-5.403 (6.5)
rs28366155	-4.47 (11.5)	-4.467 (17)	-4.946 (12)
rs743400	-4.47 (11.5)	-4.467 (17)	-4.946 (12)
Smoking pack years	-0.067 (13)	-0.067 (11)	-0.074 (15)
rs2395028	-3.731 (15)	-3.729 (15.5)	-4.129 (15)
rs2516454	-3.731 (15)	-3.729 (15.5)	-4.129 (15)
rs2516463	-3.731 (15)	-3.742 (15)	-4.129 (15)
rs2596458	-3.731 (15)	-3.729 (15.5)	-4.129 (15)
rs2596480	-3.731 (15)	-3.729 (15.5)	-4.129 (15)
Alcohol quantity	-1.104 (18)	-1.103 (18)	-1.222 (18)
rs2242660	1.421 (20)	1.454 (20)	1.632 (19)
rs1266076	1.45 (20.5)	1.449 (20.5)	
rs805301	1.45 (20.5)	1.449 (20.5)	
rs813115	1.45 (20.5)	1.449 (20.5)	
rs805302	1.431 (22.5)	1.447 (23)	
Alcohol use	-2.238 (23)	-2.236 (20)	-2.476 (22)

Note: [a] order consists of ranking the variables according to the amount of Larsen score variation explained, if outside the top 25 for that fold and that run then they are given an order equal to the total number of variables (equal last as not selected in the model). The median order is then calculated over the folds within a run and then over the runs. Hence, a median order >25 implies for some folds these variables were not selected.
Although parameter estimates (beta coefficients) are presented here for comparison across models they require careful interpretation as the size of them will depend on how correlated that variable is with other variables in the model.

5.4. Summary

Three imputation methods were explored using a subset of the 'GWAS SNPs' data (N=394) focusing on a small section of SNPs in relatively high (315 SNPs) or low (248 SNPs) LD. 'Quick' imputation and NIPALS imputation were investigated (as described in chapter 4) along with the software IMPUTE2 which uses observed haplotypes to predict missing data by comparing them against a reference set. All imputation methods performed similarly with no sizable disparity between the predictive ability or the variables selected.

To extend the 'percentage fold' method to be able to model fewer subjects, various model creation strategies were explored. These included various numbers of variables to extract (in each fold) and whether variables which were extracted in 2/5 folds or 3/5 folds in all runs would be kept for the final model. The number of variables extracted each time in the model was found to be critical, too many and the prediction ability decreases, too few and you may be missing important potential predictors of severity. The more variables selected to be extracted from the model each time, the more ended up in the final model as there was more chance for them to be selected in 2/5 folds or 3/5 folds in all runs. Therefore, it was a balance between the criteria used to retain variables for the final model and the number of variables chosen to extract each time.

It was decided to select NIPALS imputation as the imputation method for future research. This was chosen to avoid the potential under-estimation of the variation using the 'quick' imputation method (section 4.5) and would substantially reduce the time needed to impute the 'GWAS SNPs' data using the IMPUTE2 software. IMPUTE2 would also be problematic to take forward as the chip used for the GWAS selected SNPs to represent the entire genome and were not necessarily in high LD. Therefore, even after IMPUTE2 has tried to replace missing values, a further imputation method would be required to replace any which could not be imputed using IMPUTE2.

Initial reproducibility checks indicated that although there may be some variation in the lower order variables selected, which were caused by using different folds of the subjects in the runs, the top selected variables remained the same.

Using just the environmental variables and these small sets of SNPs, the final models achieved a maximum correlation between the actual and predicted Larsen score of r=0.588. This was only just higher than the correlation observed using just the first four variables of disease duration, symptom duration, age at time of diagnosis and age at onset of symptoms alone in the model (r=0.585). This suggests that none of the variables after the first four are really increasing the predictive ability of the model. Whilst there are concerns at this stage that on the reduced set of GWAS SNPs, a poor predictive ability was observed outside of the top four variables, it is hoped once the entire genome of SNPs are included, more promising results may be obtained.

6. SPLS regression of Larsen score – 'GWAS SNPs' dataset

6.1. Aims

The aim of this chapter is to:

- Incorporate the lessons used above using the 'percentage fold method' to model the 'GWAS SNPs' dataset using SPLS and see how accurately the Larsen score can be predicted.
- Investigate alternative methods of model fitting to increase efficiency such as the 'average rank' method.
- Interpret the final selected model fitted on the 'GWAS SNPs' dataset to include a comparison with univariate SNP testing.

6.2. 'Percentage fold' method on 'GWAS SNPs' dataset

6.2.1. Larsen score distribution and other measures in 'GWAS SNPs' dataset

Figure 2.1 reveals the Larsen score to have a right skewed distribution with 13.6% (137/1009) of the subjects with a Larsen score of zero. 45% of the subjects with a Larsen score of 0 are ACPA positive and have disease durations ranging from 3 to 57 years (mean=8.65, median=6.0). When the subjects in the 'GWAS SNPs' dataset are examined (N=394), the distribution is quite different. Although still right skewed, only nine subjects (2%) have a zero Larsen score (Figure 6.1). There is still a wide representation of the duration of disease although many of the subjects with long disease duration and zero Larsen score (as seen in Figure 2.2) are not included (Figure 6.2). It therefore appears likely that the subjects selected for the GWAS were not entirely at random and perhaps were selected to represent the spread of Larsen score results excluding extreme outliers. This could result in a biased sample which will need to be considered when interpreting the results.

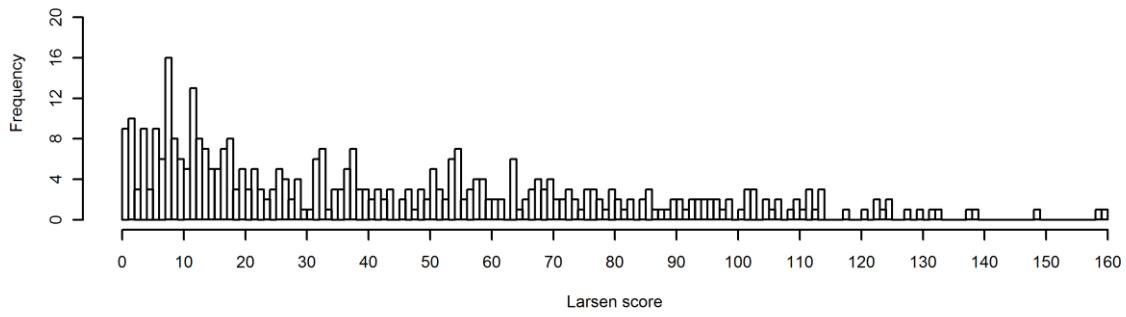


Figure 6.1 Larsen score distribution for GoRA subjects with GWAS SNPs

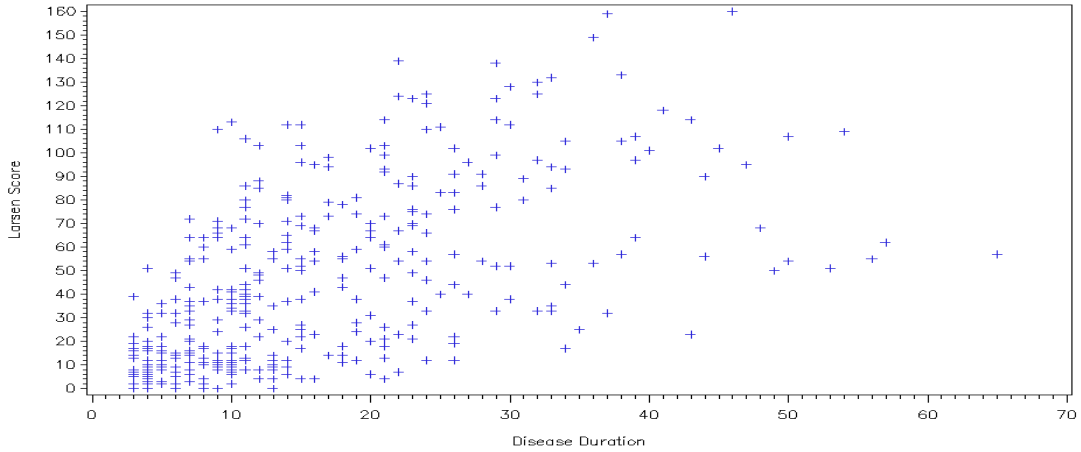


Figure 6.2 Larsen score plotted against disease duration for GoRA subjects with GWAS SNPs

Despite the notable change to the distribution of the Larsen score, the subjects in the 'GWAS SNPs' dataset are similar to the 'all subjects' dataset with respect to their demographics and environmental factors (gender, age, height, weight, BMI, waist circumference, smoking and alcohol habits). They are also similar with respect to the average and range of disease duration, symptom durations, age at disease onset and age at onset of symptoms. There is no notable change in the proportion of subjects with shared epitope alleles or with positive ACPA and RF categories. This indicates the sample is representative of the original full cohort.

6.2.2. Inclusion of additional environmental variables

Unlike the 'all subjects' dataset analysis in chapter 4, it was decided to add to the analysis three additional environmental variables (BMI, ACPA category and ACPA value). These variables were not originally included, because the measures in this cohort of patients were recorded at very different times after disease diagnosis. Therefore BMI and ACPA values may be influenced by the severity of disease and hence be confounded with the Larsen score. However, as described in section 2.4.2.3, numerous studies have shown that a low BMI leads to a higher radiographic joint damage and there is increasing evidence that ACPA categorises patients into two distinct subsets of disease (van der Helm-van Mil and Huizinga, 2008, Geng et al., 2012, Ibn Yacoub et al., 2012a, Jansen et al., 2002). In a recent meta-analysis the odds of erosions was shown to be greater for ACPA positive subjects compared to ACPA negative subjects (Taylor et al., 2013). It was therefore felt important to include these variables as predictors in the analysis. The variables representing smoking and alcohol use may also be unreliable as subjects may not have a constant usage over the course of the disease. This is a limitation of the available data, as severity and the covariates are only recorded once, which on average is recorded 14.7 years (STD=10) after disease diagnosis. Although it could be argued that RF should also be included in the modelling, it was decided against the inclusion, as RF is more variable over time than ACPA (Mjaavatten et al., 2011).

6.2.3. Modelling the data in smaller blocks

Using the R package mixOmics (González et al., 2011, Lê Cao et al., 2009) version 3.0, it was not possible to perform a SPLS regression on the 'GWAS SNPs' dataset (394 subjects with 325482 SNPs) in a single analysis due to the excessive number of variables causing R to reach its maximum memory limit in the singular value decomposition of the matrix.

Hierarchical PLS regression (Eriksson et al., 2004, Fonville et al., 2010, Vinzi et al., 2010), fits PLS models to conceptually meaningful blocks and then uses the loading vector from the lower level model to enter into a higher level model. Whilst this means that the data is modelled in smaller blocks, (enabling large datasets to be modelled), interpretation can be difficult, as you calculate higher level loading vectors from the lower level loading vectors and all variables from the initial blocks are required to form the final model. It was decided to adapt this methodology using the SPLS techniques developed in chapters 4 and 5 to perform a two stage variable selection process.

Each chromosome would form its own block and larger chromosomes would be split into two or three blocks, depending on the size of the chromosome. Chromosomes 1, 2 and 6 were split into three blocks, chromosomes 3, 4, 5 and 7 to 13 were split into two blocks and chromosomes 14 to 22, X and XY Pseudo autosomal region of X were one block each. This equated to 40 separate blocks of data with up to 12000 SNPs in each block. Increasing the block sizes above 12000 SNPs caused the models to run very slowly or the program to crash.

Unfortunately after splitting the data into blocks and completing the modelling, it was noted that the SNPS were sorted by rs number and not by their position on the chromosome. This meant that SNPs, likely to be more correlated as they are positioned closer on the chromosome, may not be modelled in the same block. It would be a very large task to sort the SNPs by position due to the format of the data in SAS®. Therefore it was decided not to repeat the analysis as the impact is hoped to be minimal due to the following reasons:

- The issue only affected chromosomes 1-13 as the rest of the chromosomes are modelled as complete blocks.
- The splitting only affected the value given to the imputed missing data which based on chapter 5 didn't appear to affect the model.
- Important correlated variables in separate blocks should be both carried forward to the higher level model although this may miss interactions.
- Even if the variables could be sorted by position before cutting off into blocks, SNPs close to the cut offs would still not have the SNPs next to them in the same block.

The problem of modelling the whole chromosome in one block is only a problem in R. Sections 6.4.4 and 7.2 explore the use of COREXpress (Magidson, 2011) and SIMCA (Eriksson et al., 2006a) which are both able to model much larger datasets. Section 7.4 supports that the above assumption was not a problem as similar results were obtained by modelling all SNPs in one block using SIMCA.

6.2.4. 'Percentage fold' method model fitting strategy

Each of the 40 blocks were treated as an individual set of data, merged with the environmental variables and missing data imputed using the NIPALS algorithm (as described in sections 4.2.2 and 5.4). The full model fitting process using the 'percentage fold' method was completed similar to 4.2.6 but with the 'GWAS SNPs' dataset changes (5-fold CV and 2/5 folds in all 50 runs) applied as described in section 5.2.1 and 5.3.2. The variables of highest predictive importance were extracted from each block. Once the top variables from each of the 40 blocks were extracted they were combined together to form a higher level model. The entire model fitting process was repeated on the reduced set of data and the most predictive variables extracted from across the entire genome.

The above approach had not been identified in the literature, however, it was decided to investigate it as an alternative approach to pre-filtering of SNPs prior to SPLS analysis (Le Cao et al., 2011).

6.3. 'Percentage fold' method results on 'GWAS SNPs' dataset

6.3.1. 'Percentage fold' method results from lower level modelling

Of the 40 blocks, the maximum number of variables selected to extract for a block were 200 and the minimum was just 10. The mean (82.25) and median (82.50) were similar and the STD was 41.34. Graphs for each of the 40 blocks along with justification of the choice of the number of variables to extract is provided in Appendix E.

For the first five blocks of data, two components were initially explored (in accordance with section 4.2.6). Similar to all previous models on the Larsen score (in chapters 4 and 5), all of the five blocks concluded that only one component was required and the whole model fitting strategy redone for each of the five blocks using just one component. As 40 blocks have to undergo this process and this stage is only used to select the number of variables to take through to the higher level model, it was decided to assume that only one component was required for all 40 models. As the first component explains the largest amount of Y variation and all models to date indicated only one component was required, this was considered a low risk time saving strategy.

6.3.1.1. Environmental variables selected

For the environmental variables, a reassuring consistency was observed when the 40 blocks of SNP data were fitted and the most important variables for predicting Larsen score extracted. The top four variables selected in the modelling of every block of SNPs were; disease duration, symptom duration, age at time of diagnosis and age at onset of symptoms respectively (Table 6.1). BMI was selected as the top 5th variable in every block of SNPs with the exception of chromosome 2 part 3, chromosome 4 part 2 and chromosome 6 part 1, where it was selected 7th, 6th and 6th respectively. Therefore, it has an average selection order of 5.1. The remaining environmental variables which were extracted in 2/5 folds in all 50 runs in at least one block of SNPs showed more variation as to whether they were selected in all blocks (Table 6.1). Whilst they were not identified in all blocks as important predictors of the Larsen score, all of these variables will be carried forward for the higher level model, as they could always be removed at a later stage if they ranked sufficiently low. Neither gender nor any of the shared epitope variables were selected in any of the models and therefore were not taken forward to the higher level modelling.

Table 6.1 Selected environmental variables from the 'GWAS SNPs' dataset Larsen score modelling

Variables	Average order of selection	% blocks extracted in 2/5 folds in 50 runs	% blocks extracted in 5/5 folds in 50 runs
Disease duration	1	100	100
Symptom duration	2	100	100
Age at time of diagnosis	3	100	100
Age at onset of symptoms	4	100	100
BMI	5.1	100	100
Smoking status	16	5	5
ACAP value	265	90	85
Smoking duration	278.11	90	85
Smoking pack years	493.84	85	75
ACPA category	2104.3	25	12.5
Alcohol quantity	2845.31	65	20
Alcohol use	3029.13	20	5

6.3.1.2. SNPs selected

SNPs being selected by the model are presented by chromosome blocks in Table 6.2. The table contains the number of SNPs requested to be extracted for each of the blocks in each fold of each run, those SNPs extracted in 5/5 folds in all 50 runs and the additional SNPs extracted in at least 2/5 folds in all 50 runs. The latter column represents SNPs which are more variable and are sometimes not selected when the patient mix changes in the folds.

The downside of using an arbitrary cut off for variables being extracted is that, after the top variables extracted have been ranked, the remaining received equal last rank. Therefore, suppose not enough variables were being extracted and a SNP was consistently just outside of the extracted number, they would not be selected for the final model. However, as a generous threshold is being used, many false positives should be carried through (variables which are thought to be predictive of the Larsen score, but they are actually not predictive). Hence, those variables identified outside of the extracted set should just explain noise in the Larsen score and not be true predictors. Removing this cut off entirely from the modelling methods is explored further in section 6.4.1.

As the SNPs selected in 5/5 folds in all 50 runs are clearly more robust than those selected in ≥2/5 folds, it was decided to run two higher level models. A minimum SNP model (with just those selected in 5/5 folds) and a maximum SNP model (with those selected in ≥2/5 folds).

Table 6.2 Selected SNPs from GWAS Larsen score modelling

Chromosome (Number SNPs selected to be extracted)	Number and identification of SNPs extracted in 5/5 folds in all 50 runs. (SNPs entered in the minimum SNP model)	Number and identification of the SNPs with ≥2/5 folds but not 5/5 folds in all 50 runs	N SNPs in maximum SNP model
1, part 1 (45)	0: None	3: rs1888182, rs998459, rs954916	3
1, part 2 (130)	23: rs2646852, rs2843157, rs6427859, rs2148322, rs2643891, rs4568808, rs6674079, rs2376907, rs4658015, rs3827707, rs4658547, rs2378494, rs2365270, rs2476020, rs2493272, rs4657226, rs4927072, rs4658518, rs6660197, rs4418557, rs2998676, rs3843257, rs2225999	19: rs3008588, rs4654552, rs4233084, rs4446947, rs2684866, rs2878284, rs4304627, rs2292096, rs4970551, rs6429673, rs6577539, rs6663404, rs2878076, rs2923905, rs3737686, rs4626924, rs3765968, rs4509550, rs6673009	42
1, part 3 (85)	11: rs10925085, rs7528217, rs10495071, rs10922014, rs12134313, rs10495064, rs11802794, rs10495072, rs12145838, rs10493308, rs7543680	8: rs10926974, rs12079354, rs7524397, rs12124014, rs12758038, rs10922200, rs12080578, rs12745508	19
2, part 1 (200)	38: rs1374153, rs1013940, rs955038, rs1250043, rs1862893, rs1467219, rs2028863, rs1025080, rs1370666, rs2084779, rs749460, rs1996634, rs16909, rs1882449, rs955799, rs1867801, rs1921814, rs2053761, rs974813, rs1367415, rs1438065, rs2044302, rs1224540, rs908275, rs1574181, rs929633, rs1899013, rs1367457, rs2114619, rs1514687, rs867458, rs1357182, rs1453496, rs985091, rs893811, rs893451, rs1526638, rs877328	25: rs1377389, rs163507, rs1317981, rs2079538, rs981602, rs1383413, rs1568378, rs2110745, rs960615, rs1035833, rs1432216, rs1978368, rs355810, rs1146025, rs1470494, rs2080711, rs959257, rs1006413, rs1921772, rs905968, rs917237, rs295145, rs2052928, rs288057, rs882672	63
2, part 2 (100)	16: rs2218495, rs2680182, rs2889450, rs4851042, rs4668412, rs2304673, rs2276630, rs4389306, rs2883605, rs2215801, rs6543316, rs4671614, rs6723319, rs3771688, rs4850546, rs4673304	14: rs2515402, rs2592274, rs3755427, rs2305491, rs2887311, rs2894500, rs6725634, rs4669075, rs6727113, rs6738271, rs2292869, rs2366812, rs4073806, rs4402808	30
2, part 3 (80)	12: rs11211654, rs9326161, rs11897977, rs10497610, rs11683588, rs11902332, rs12992554, rs10928155, rs7581781, rs7602332, rs6758704, rs10931024	2: rs16830673, rs10201889	14
3, part 1 (40)	3: rs1558910, rs1532190, rs1950091	2: rs1482601, rs925680	5
3, part 2 (110)	13: rs6788054, rs9841471, rs9847705, rs6762974, rs6806453, rs9842509, rs7651268, rs11131004, rs9290621, rs7633462, rs10804755, rs9853204, rs9823801	13: rs10513603, rs11718444, rs13082763, rs9833136, rs7630937, rs6763272, rs7622674, rs9837726, rs7609860, rs7629690, rs9863209, rs7643807, rs13078878	26
4, part 1 (90)	16: rs1449779, rs2015829, rs1375775, rs1449760, rs2101201, rs2242310, rs4391009, rs3851433, rs3970313, rs4339259, rs971079, rs1399531, rs1375738, rs2167955, rs4689290, rs1565534	8: rs1320101, rs1514951, rs1850756, rs1107679, rs2014472, rs2114148, rs243970, rs260122	24
4, part 2 (140)	22: rs7661111, rs9884240, rs6814036, rs10031745, rs6856061, rs9291499, rs9291501, rs7663212, rs12500896, rs7664301, rs6828208, rs4833470, rs4696891, rs6846226, rs6535700, rs17088003, rs10017867, rs10516659, rs12512924, rs13142037, rs6811964, rs6537420	21: rs10517429, rs6552386, rs13105340, rs10517960, rs13108614, rs10006264, rs7666497, rs10012647, rs10029929, rs11097585, rs12641938, rs12642793, rs12645705, rs13106777, rs13121783, rs17006928, rs4694421, rs7661337, rs10005962, rs4865293, rs7684707	43
5, part 1 (120)	13: rs2933639, rs346650, rs265005, rs309556, rs784480, rs2635636, rs993764, rs1296135, rs262029, rs1992707, rs712572, rs36951, rs37181	22: rs26821, rs346660, rs2560623, rs183752, rs2591961, rs695103, rs784482, rs890294, rs2277054, rs245674, rs36676, rs721834, rs832575, rs923957, rs2968014, rs346658, rs1013429, rs1465291, rs1499764, rs187580, rs526231, rs674726	35
5, part 2 (10)	None	None	0
6, part 1 (90)	21: rs760609, rs1323056, rs2178899, rs2023075, rs220704, rs1042663, rs1046080, rs3065, rs497239, rs547154, rs550605, rs194679, rs785144, rs1047033, rs1407220, rs2181437, rs511294, rs679013, rs438999, rs154986, rs341103	34: rs927407, rs541862, rs714051, rs879036, rs154980, rs1984814, rs991760, rs1029295, rs1116029, rs1497737, rs604683, rs2181351, rs941816	34

Chromosome (Number SNPs selected to be extracted)	Number and identification of SNPs extracted in 5/5 folds in all 50 runs. (SNPs entered in the minimum SNP model)	Number and identification of the SNPs with ≥2/5 folds but not 5/5 folds in all 50 runs	N SNPs in maximum SNP model
6, part 2 (110)	15: rs3115667, rs2272593, rs2395672, rs2844586, rs4454114, rs4714801, rs6904223, rs6931700, rs3130617, rs2281272, rs2297308, rs2857597, rs2523619, rs3117133, rs7745708	27: rs6916278, rs2272450, rs6921518, rs2516463, rs3129867, rs4140408, rs6938822, rs4896481, rs6913013, rs3777536, rs4714942, rs7739145, rs2327832, rs2894249, rs3129941, rs6939790, rs7741687, rs2327212, rs4335008, rs3129871, rs4946442, rs3129907, rs4895851, rs2505951, rs3823034, rs7752020, rs6457689	42
6, part 3 (65)	8: rs12211935, rs7770227, rs10498671, rs10484566, rs10457277, rs9372396, rs9380681, rs17673852	10: rs28366162, rs9346693, rs10434890, rs10080898, rs10214886, rs9485215, rs9469300, rs10046277, rs9350822, rs9469312	18
7, part 1 (95)	18: rs2132046, rs1843933, rs4099560, rs3847015, rs705321, rs2727805, rs1878808, rs2075756, rs819454, rs2065668, rs702479, rs2245192, rs3211834, rs259152, rs2024365, rs2711481, rs41700, rs2286680	12: rs36878, rs3801279, rs855686, rs2372052, rs2402061, rs2527636, rs926201, rs1024542, rs1819819, rs2528528, rs855679, rs3779107	30
7, part 2 (75)	10: rs6597458, rs6459830, rs10499471, rs17625938, rs6947058, rs12539883, rs13238018, rs10279978, rs10499472, rs7789085	11: rs11763025, rs17558927, rs6948116, rs11767604, rs4722166, rs6465211, rs7779140, rs7805617, rs7808226, rs6462473, rs12718939	21
8, part 1 (95)	4: rs1354969, rs3943520, rs4551310, rs1161534	4: rs2941647, rs2957422, rs311390, rs366276	8
8, part 2 (75)	1: rs7816057	8: rs13254276, rs7386980, rs7386656, rs10096683, rs10504242, rs10091402, rs7461129	8
9, part 1 (90)	7: rs1754068, rs513806, rs1341063, rs1777045, rs3811159, rs944638, rs2153240	17: rs783455, rs1407808, rs2777877, rs1668978, rs2275003, rs967671, rs1887890, rs3016756, rs7776, rs2149171, rs2164001, rs3750433, rs488948, rs2210369, rs1339490, rs914842, rs2147263	24
9, part 2 (110)	13: rs10511820, rs7873559, rs12001157, rs7864699, rs4842019, rs4743316, rs10868794, rs5012630, rs7848626, rs4074426, rs4842064, rs11793528, rs4742515	11: rs4146765, rs4741199, rs10758631, rs10976375, rs7869933, rs4842153, rs7025731, rs10441723, rs10115893, rs12336000, rs10810351	24
10, part 1 (45)	4: rs1705013, rs1860404, rs4457655, rs1111267	3: rs2115819, rs3763745, rs884147	7
10, part 2 (70)	10: rs11016877, rs7477944, rs10509236, rs11593905, rs6584162, rs10905618, rs11017878, rs6479789, rs6481891, rs7897741	7: rs10490907, rs6585536, rs10509958, rs7905689, rs12262693, rs17645752, rs7898685	17
11, part 1 (95)	19: rs470215, rs470747, rs508835, rs470168, rs2429862, rs2499937, rs1005511, rs2303973, rs2063024, rs1237999, rs2445296, rs541458, rs1783229, rs677909, rs598373, rs903514, rs947837, rs487728, rs2472528	14: rs1848082, rs725103, rs543293, rs2472527, rs756852, rs874548, rs964720, rs1440718, rs953816, rs1600591, rs2450411, rs2875379, rs475639, rs685320	33
11, part 2 (160)	33: rs4609584, rs7948160, rs10792830, rs10898438, rs10835941, rs4623925, rs4963214, rs16910726, rs7111383, rs4963212, rs7948050, rs4758331, rs3851179, rs10500616, rs10791200, rs11821612, rs3858451, rs4758322, rs10790866, rs3819100, rs7102738, rs7121743, rs7925573, rs4512811, rs4581433, rs7924850, rs7483826, rs4754673, rs4529888, rs7926469, rs4938097, rs7941509, rs7103780	22: rs7110845, rs7931095, rs10830200, rs7935178, rs10767936, rs12421053, rs12807255, rs3782115, rs4294557, rs4480535, rs4938377, rs4938800, rs7130116, rs11224629, rs11232234, rs7112940, rs11019392, rs11019435, rs2943510, rs3855349, rs4121403, rs7117594	55
12, part 1 (170)	35: rs2686386, rs3751143, rs1879390, rs516505, rs4523751, rs1861918, rs526058, rs328765, rs757354, rs1496858, rs4129599, rs1434725, rs1544608, rs1517727, rs741628, rs1820460, rs844066, rs270881, rs1304341, rs3934768, rs2117322, rs328759, rs4246260, rs1488144, rs2366796, rs871257, rs2158091, rs937529, rs3993375, rs1894791, rs1562729, rs2971589, rs1920438, rs2700568, rs278899	53: rs4131751, rs772700, rs917915, rs1027569, rs1569020, rs2304274, rs4237941, rs954147, rs1093291, rs1400138, rs1839402, rs1851094, rs979678, rs1426437, rs1471132, rs1631980, rs1669921, rs1816854, rs2398526, rs3782614, rs445467, rs962051, rs1385374, rs1513047, rs1725789, rs1873347, rs303784, rs1032332, rs1148985, rs1566514, rs1682593, rs1800973, rs1907087, rs1968964, rs2111177, rs2320501, rs2870951, rs4502065, rs472100, rs722097, rs730165, rs759518, rs1847459	78
12, part 2 (90)	11: rs10878920, rs4760805, rs10505938, rs6489900, rs11177669, rs4764112, rs7309856, rs11106529, rs11104713, rs7312143, rs10842797	13: rs4761777, rs7972233, rs7314666, rs10506049, rs7959196, rs17523988, rs11169992, rs11170524, rs5019656, rs10842223, rs11104703, rs4765138, rs6582065	24

Chromosome (Number SNPs selected to be extracted)	Number and identification of SNPs extracted in 5/5 folds in all 50 runs. (SNPs entered in the minimum SNP model)	Number and identification of the SNPs with ≥2/5 folds but not 5/5 folds in all 50 runs	N SNPs in maximum SNP model
13, part 1 (85)	12: rs4770403, rs892533, rs1413882, rs572761, rs795659, rs1159207, rs1056114, rs2091337, rs1945503, rs744574, rs912656, rs1324978	12: rs837309, rs4612929, rs1591478, rs837344, rs560873, rs269588, rs728993, rs735721, rs1341476, rs1354837, rs280392, rs1927837	24
13, part 2 (65)	9: rs9563202, rs6420308, rs7319076, rs9571979, rs9588434, rs9300986, rs9563119, rs12866357, rs9585457	4: rs7982644, rs9525291, rs7999681, rs9301172	13
14 (90)	13: rs4898652, rs7142677, rs7157967, rs10873293, rs8004595, rs10483876, rs1652593, rs4601978, rs10484017, rs8013529, rs11622224, rs3939209, rs1479748	6: rs2215590, rs12890068, rs2049826, rs4901639, rs10483877, rs7150135	19
15 (80)	14: rs782933, rs8032023, rs16939900, rs2036534, rs7175069, rs2133570, rs4534820, rs4589506, rs1356782, rs2133127, rs2241494, rs890158, rs2871886, rs4887053	14: rs4924057, rs4506837, rs1584407, rs1975242, rs9944198, rs768546, rs2852078, rs650716, rs570763, rs6495309, rs6495894, rs1356779, rs2682911, rs8035668	28
16 (30)	0: None	4: rs1834037, rs4473203, rs4522429, rs916768	4
17 (55)	9: rs879606, rs799923, rs4795369, rs7208487, rs2777899, rs2941503, rs2941504, rs12453682, rs1292034	5: rs2061342, rs4074770, rs4793832, rs7215464, rs3785982	14
18 (30)	0: None	None	0
19 (60)	3: rs3745333, rs11671924, rs10420734	7: rs11086047, rs7252814, rs10411465, rs1141371, rs2283575, rs8108252, rs8111930	10
20 (50)	5: rs11299, rs6050732, rs6046528, rs2747405, rs6044003	3: rs285164, rs6050709, rs373561	8
21 (120)	24: rs2299742, rs383700, rs8132953, rs232456, rs2839315, rs440666, rs7277065, rs2826825, rs2154420, rs2822430, rs2833845, rs7280944, rs2837801, rs2833886, rs232518, rs2236436, rs2822429, rs2827308, rs2834049, rs2833629, rs722682, rs2834157, rs2837868, rs762417	18: rs2178832, rs232496, rs1910635, rs468192, rs1787438, rs2027715, rs2838679, rs2824693, rs2832451, rs2833836, rs2837985, rs2850163, rs2839112, rs400603, rs1554936, rs2837716, rs463117, rs7280236	42
22 (35)	6: rs1076933, rs2017317, rs2072711, rs2018293, rs5993935, rs5998876	1: rs5997898	7
X chromosome (45)	15: rs5928558, rs2266806, rs4828734, rs6520724, rs10522027, rs1926105, rs6630822, rs5990454, rs12852732, rs4573446, rs5920765, rs3132267, rs578264, rs4826799, rs2172209	1: rs5955786	16
XY Pseudo autosomal region of X (25)	7: rs5941380, rs10127367, rs4074621, rs35047434, rs306910, rs34438890, rs2750171	0: None	7
Total number SNPS	493	426	919

NOTE: Blocks 7 part 2, 14 and X amended the criteria for a SNP to be fitted in the model from <92% all 0-0.5 to be <85% to enable the model to fit. Whilst data was available for the Y chromosome and the mitochondrial region, there was not enough data present with enough variation to enable any SNPs to be modelled.

6.3.2. 'Percentage fold' method results from higher level modelling

As shown in Table 6.2, two models would be taken forward; firstly the 'minimum SNP model' which consisted of SNPs extracted in 5/5 folds in all 50 runs (González et al., 2011, Magidson, 2011) and secondly the 'maximum SNP model' which consisted of SNPs extracted in ≥2/5 folds in all 50 runs. The same environmental variables were used in both models. If there was no benefit in adding the extra SNPs, then the process could be amended in future to only retain those SNPs selected in all folds. Model fitting using the 'percentage fold' method was the same as described for each of the blocks in the lower level modelling (section 6.2.4).

6.3.2.1. Minimum SNP model

493 SNPs and 12 environmental variables formed the higher level minimum SNP model. All variables were entered into the model and then in turn, the top five to 500 variables were extracted calculating each time the R^2-CV using 5-fold CV on the 394 subjects. It was hoped that this would display the optimum number of variables for the model, whilst preventing over fitting, as the R^2 is calculated under CV. Unlike modelling the smaller number of SNPs in chapter 4 (which could never achieve an R^2-CV greater than around 0.6), Figure 6.3 shows that using the selected SNPs from the lower level modelling, the R^2-CV continued to increase, the more variables that were included, up to a value of 0.9316 at 500 variables. Therefore all 505 variables were kept for the final minimum SNP model. Further investigation can be found in section 8.3.2 into the over fitting of these models.

Performing a standard PLS model (No variable selection or CV) resulted in a correlation between the actual Larsen score and the predicted Larsen score of 0.95 (Figure 6.4). The mean absolute difference between the actual and predicted Larsen score was 8.90 (median=7.36, STD=6.99, min=0, max=47.35). This was a big improvement compared to modelling a smaller set of SNPs but on a larger number of subjects in chapter 4.

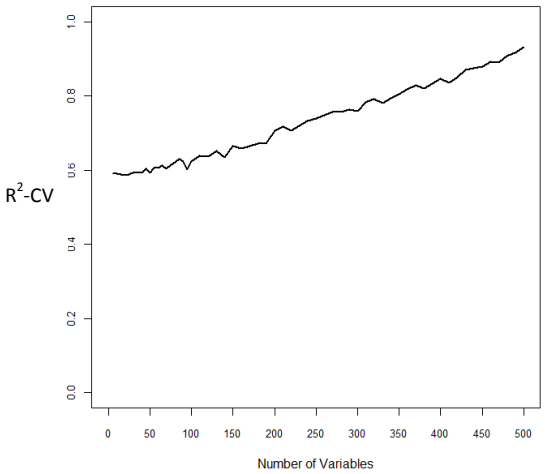


Figure 6.3 Minimum SNP model R^2-CV for each number of variables extracted

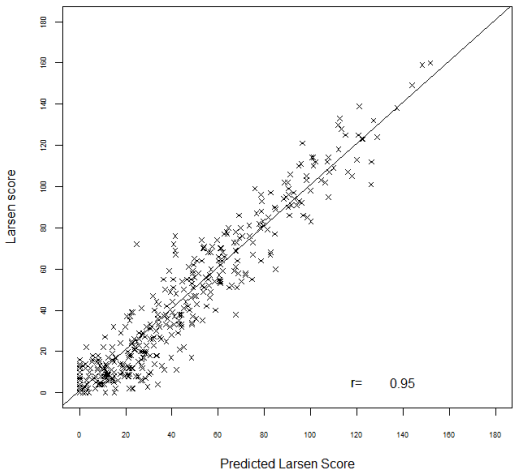


Figure 6.4 Minimum SNP model actual Larsen score versus predicted Larsen score

Using the standard mixOmics 'valid' function version 3.0 (González et al., 2011, Lê Cao et al., 2009), a PLS model using 5-fold CV (keeping all 505 X variables) was used to estimate anticipated model performance on independent dataset. This resulted in an R^2 of 0.937 and a R^2-CV of 0.874. However, as the original selection of variables was performed under multiple CV models, it is no surprise that the model performs well under CV. Therefore, this is not a reliable method to use to estimate how this model would perform on an independent set of data. Further validation techniques are explored in chapter 8.

6.3.2.2. Maximum SNP model

919 SNPs formed the maximum SNP model, having met the criteria in the lower level models of being selected in at least 2/5 folds in all 50 runs. Figure 6.5 shows the R^2-CV calculated using 5-fold CV on the 394 subjects extracting from five to 900 variables. The maximum R^2-CV was obtained at 0.942 with the full 920 variables and is only just higher than the 0.9316 obtained using the minimum SNP model.

Performing a standard PLS model (No variable selection or CV) resulted in a correlation between the actual Larsen score and the predicted Larsen score of 0.963 (Figure 6.6). This is slightly higher than the correlation of 0.95 which was obtained using the minimum SNP model. The mean absolute difference (7.86) between the actual and predicted Larsen score was also slightly better as were the other summary statistics (median=6.80, STD=5.92, min=0, max=30.30).

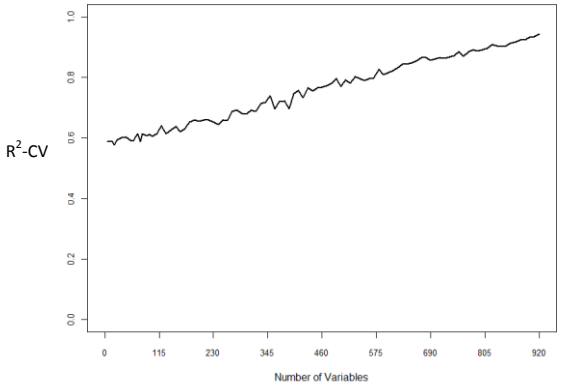


Figure 6.5 Maximum SNP model R^2-CV value for each number of variables extracted

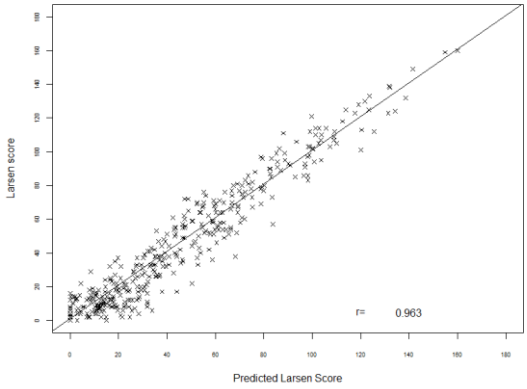


Figure 6.6 Maximum SNP model actual Larsen score versus predicted Larsen score

The maximum SNP model was run under 5-fold CV which resulted in an R^2 of 0.949 and a R^2-CV of 0.894 compared to the previous R^2 of 0.937 and a R^2-CV of 0.874 under the minimum SNP model.

Therefore the increase by adding an additional 426 variables was only very slight. Whilst there is a good decrease in the maximum prediction error between actual Larsen score and predicted Larsen score (47.35 to just 30.30), the increase in accuracy is quite likely to be due to over fitting of the model (as described in section 3.7).

6.4. Investigation into alternative modelling strategies

The process used for model fitting so far was quite laborious and time consuming, due to the number of manual steps and separate runs of CV models. Full model fitting on the 40 blocks of the 'GWAS SNPs' dataset took over one week to run. The following section investigates ways to speed up the process and looks into alternatives to using R software and the package mixOmics. The changes made to the mixOmics version 3.0 'valid' and 'spls' functions for the analysis in this section are shown in Appendix D. In addition, the code to produce the model fitting strategy after the amendments discussed in sections 6.4.1 and 6.4.2 have been made, is shown in Appendix F.

6.4.1. 'Average rank' method compared to 'percentage fold' method

Previously, using the 'percentage fold' method, the number of variables to extract for each block of SNPs was investigated by extracting; 1, 2, 3, 4, 5 to 100 by 5 and 110 to 250 by 10 variables. This equates to 39 models for each of the 40 chromosome blocks, resulting in 1560 runs of 5-fold CV. Once the number of variables to extract was selected for each of the 40 blocks, the model had to be re-run another 50 times for each of the 40 chromosome blocks under 5-fold CV extracting the chosen number. Variables were then ranked for each fold and each run and those not in the top set extracted were given a score of equal last rank for that model. The rank was then averaged (using the median) over the folds and then over the runs but not used further to select the variables. Variables were only taken forward to the higher level model if they were extracted in at least 2/5 folds (or 5/5 folds) in all 50 runs.

Section 5.3.2 has already revealed considerable variation in the variables selected, when the number chosen to extract and the number of folds a variable has to be selected (i.e. 2/5 or 3/5), are changed. It was also apparent that although the model calculates ranks of all of the variables, this rank is not being used to its full potential, as those not in the top extracted are ranked equal last. There was concern that rare genotypes, which may not have sufficient data to be selected as important in all 5 folds, would not be given an opportunity to be selected for the final model. Therefore, in order to automate the process, reduce the number of models needed to be run and explore if rarer SNPs could be better selected, a new approach entitled the 'average rank' method was investigated.

SPLS models were fitted ranking all variables by the absolute size of each variable's loading. The median of this rank was calculated over the 5-folds of CV and then across the 50 runs. After all models were fitted, if a variable was ranked less than 200 on average, it was retained for the higher level model. The value 200 was selected as it ensured that less than 8000 variables would go through to the higher level model (200 x 40 blocks of SNPs) and it was the maximum number selected to extract in any of the SNP blocks when they were investigated individually in section 6.3.

It was immediately apparent that this new method was a fairer method to identify potentially important rarer SNPs. Figure 6.7 demonstrates for the first part of chromosome 2, that rs2081776 is only selected in the top 200 for 60 of the 250 models (5 folds * 50 runs). However on average it was selected as the 56.5th SNP. For the other runs, if it had been selected at all, it would have had a higher median rank. Therefore, it must have been excluded from the other folds due to low frequency of the minor allele. This variable would not have met the criteria of being in ≥2/5 folds in all 50 runs, however it did meet the average rank <200 criteria. The new method allowed for rarer SNPs to still have a chance of being selected for the final model.

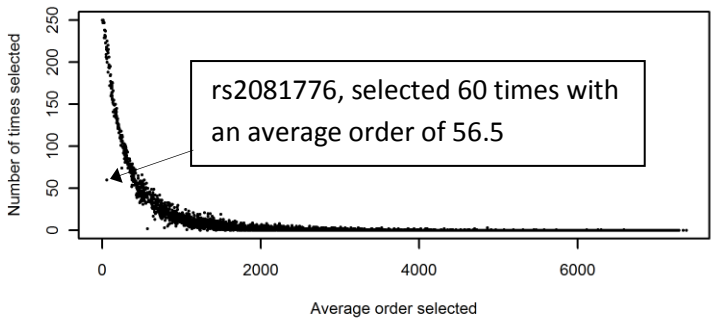


Figure 6.7 Chromosome 2: part 1, variable selection under 50 runs of 5-fold CV

The 'average rank' method created a comma separated variable file of the variable name, number of times it was selected and the average order of selection. These files for each block of SNPs could be automatically read back into R, the variables ranked in the top 200 retained and entered into the higher level model without any manual intervention. The process was therefore substantially quicker than investigating the number of variables to extract for each block by studying it graphically and manually entering the number.

3766 variables were selected using this method from the lower level models to enter into the higher level model. The same process was then used for the higher level model using 50 runs of 5-fold CV and calculating the median order of selection (rank). Rather than having to impose a restriction on this final model for a variable to be retained, all variables are ranked, which enables investigation into how much the model is improved by the addition of including subsequent variables into the model.

The 'percentage fold' method shown in section 6.3.2.1 created a higher level minimum SNP model containing 493 SNPs and 12 environmental variables. The model could predict the Larsen score with an R^2 of 0.937, a R^2-CV of 0.874 and a correlation between actual and predicted Larsen score of r=0.95. Using the 'average rank' method and the top 505 variables to match the number of variables used in the 'percentage fold' method, the predictive ability and the variables selected were compared. The 'average rank' method achieved similar predictive ability to the 'percentage fold' method; R^2=0.940, R^2-CV =0.878 and the correlation between the actual and predicted Larsen score was 0.953 (Figure 6.8). Comparing the variables selected; 353 of the 505 were the same variables. 144 new variables were selected which were not in the original minimum data model. These tended to be variables which were selected infrequently; however, when they were selected, they were in the top ranked variables. They would have not been selected previously due to the criteria that they have to be selected in 5/5 folds in all 50 runs. As these could be important but rarer SNPs, this automated method using variable ranks was considered the better method.

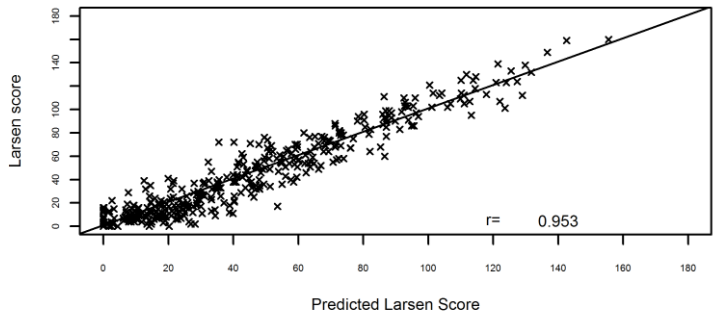


Figure 6.8 Actual Larsen score versus predicted Larsen score for higher level model (505 variables)

6.4.2. Reducing the number of runs required using the 'average rank' method

Although the use of the 'average rank' method removed the need for manual intervention and so could be run 24 hours a day, the modelling process still took approximately 50 hours on a standard computer. It was therefore decided to investigate whether using 50 runs could be reduced. This reduced the running time from 50 hours to approximately 8 hours on a standard computer. Using just 10 runs, the correlation between actual and predicted Larsen score was the same 0.953 (Figure 6.9). The R^2 =0.942 and R^2-CV= 0.882 based on a model containing 505 variables were actually slightly improved but this is likely to be due to random variation.

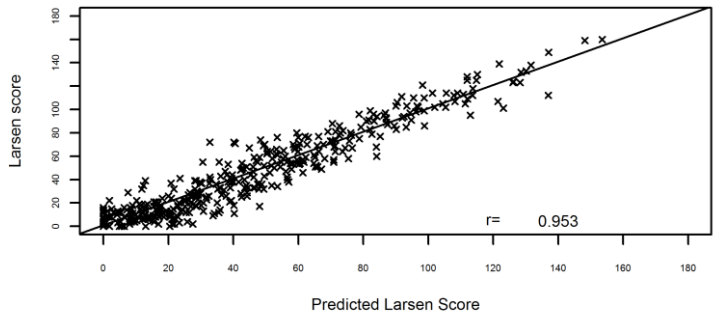


Figure 6.9 Actual Larsen score versus predicted Larsen score for 10 run higher level model (505 variables)

456 of the 505 variables were in both the 50-fold and 10-fold models. This suggests it may not be worth the extra 40 runs of the 5-fold CV on each block of data. With just 10 runs the model fit is approximately the same with very similar variables selected. Further exploration into reducing the 10 runs to just one run (which runs in approximately one hour) is described in section 7.3.

Using the median rank of the size of each variable's loading to determine the importance of the variable, appears a better method than using the number of times a variable was extracted, because the latter is dependent on the number of variables you choose to extract. The 'average rank' method allows identification of rarer but potentially important SNPs.

6.4.3. Using Bootstrapping instead of CV

CV is traditionally used for model selection, as it estimates the expected predicted error forming models trained on one set of subjects and tested on another (Hastie et al., 2009, Chapter 7.10). The 'average rank' method detailed in section 6.4.1 runs multiple CV models to select variables to take through to the higher level modelling, based on the size of the PLS loading for each variable. The CV predicted error is not used on any of the lower level modelling. However, once the higher level model is fitted the CV prediction error is used to assess how well the model performs. Not surprisingly, it appears to perform well as the variables were selected from the lower level models because they are important in the prediction consistently across multiple CV models.

An alternative method of forming a number of new sets of the data is called bootstrapping. Each new set is formed by sampling subjects randomly with replacement, so that each set are of equal size to the original number of patients.

It was decided to create 100 bootstrapped sets of patients (sampling 394 subjects from the original sample) with replacement for each of the 40 blocks of SNPs. The 'average rank' method described in section 6.4.1 was applied to the 4000 datasets after the removal of any SNPs which had more than 92% all of genotype 0. Per the 'average rank' method, variables were ranked according to the size of their PLS loading and the median rank over the 100 samples calculated for each of the 40 blocks. Any variables with a median order of selection of <200 were carried forward to the higher level model. The 'average rank' method was then applied to the higher level model using a further 100 bootstrapped samples. Variables were ordered by the median rank across the 100 models. In order to compare the performance with section 6.4.2, 505 variables were retained for the final model.

The correlation between actual and predicted Larsen score from the higher level model was 0.936 (Figure 6.10). The R^2=0.926 and R^2-CV=0.855. All performance estimates were slightly lower than the model formed using multiple CV models (r=0.953, R^2=0.942 and the R^2-CV=0.882 using the method in 6.4.2), however, this could be due to chance.

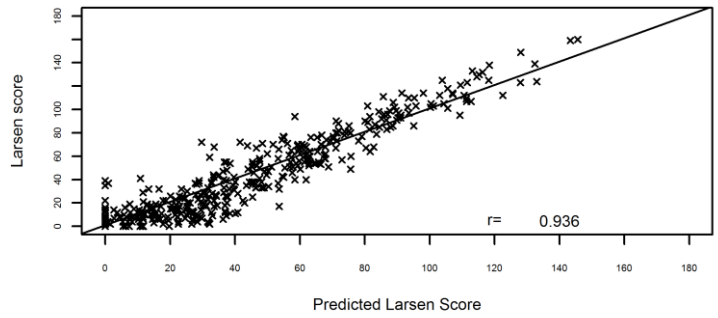


Figure 6.10 Actual Larsen score versus predicted Larsen score for bootstrapped higher level model (505 variables)

For simplicity, just the top 100 variables were compared across the 'average rank' method using 10 runs of 5-fold CV compared to bootstrapping. The same five environmental variables were selected; disease duration, symptom duration, age at time of diagnosis, age at onset of symptoms and BMI. Of the 95 SNPs which both models selected to be in the top 100, 70 were identical leaving

just 25 differences shown in Table 6.3. Whilst it would be expected that these 25 appear lower down the ranked lists of SNPs, this was not the case. Six SNPs were different in the top 30. The six SNPs chosen by the bootstrapping method in the top 30 which were not in the 10 runs of 5-fold CV were; rs10488483, rs192214, rs10512270, rs10488482, rs1514920 and rs7954187 and all had no subjects with a genotype =2. They also had a relatively small sample size with genotype=1 (N=51, 24, 21, 34, 28, 27) respectively. In contrast, the six chosen by the 10 runs of 5-fold CV, all had some subjects with genotype =2. The 10 runs of 5-fold CV did select one SNP in the top 30 with no subjects having genotype=2, however, this was also selected in the bootstrapped sample. The other 19 SNPs selected by the bootstrapping method but not by the 10 runs of 5-fold CV were spread evenly throughout the top 95, with no clear reason for not being selected in both models. 18 of the 25 were statistically significant at the 5% alpha level using a univariate NB model as described in section 4.2.5. This indicated that the bootstrapping method appeared to identify more SNPs which were significant at the 5% level when tested univariately (Table 6.3).

Table 6.3 List of the 25 SNPs selected in the Bootstrapping model not in the 10 run 5-fold CV model

SNP	Genotype: N subjects : Median Larsen score	NB p-value
rs10488483	0: 343: 32, 1: 51: 65, 2: NA: NA	0.0038
rs192214	0: 370: 35, 1: 24: 74, 2: NA: NA	0.0950
rs10512270	0: 373: 33, 1: 21: 77, 2: NA: NA	0.1173
rs10488482	0: 360: 33, 1: 34: 64.5, 2: NA: NA	0.0213
rs1514920	0: 366: 38, 1: 28: 12.5, 2: NA: NA	2.26E-05
rs7954187	0: 367: 38, 1: 27: 13, 2: NA: NA	5.84E-05
rs17832312	0: 357: 33, 1: 37: 64, 2: NA: NA	0.0161
rs11936270	0: 371: 35, 1: 21: 62, 2: 2: 137	0.0800
rs638383	0: 370: 38, 1: 24: 12.5, 2: NA: NA	0.0045
rs4760805	0: 264: 32, 1: 113: 47, 2: 17: 55	0.0512
rs879986	0: 367: 35, 1: 27: 64, 2: NA: NA	0.0322
rs6043954	0: 363: 38, 1: 30: 13.5, 2: 1: 32	0.0014
rs3943520	0: 152: 32, 1: 179: 37, 2: 63: 56	0.0062
rs9841471	0: 279: 43, 1: 106: 22, 2: 9: 21	5.92E-05
rs2941647	0: 239: 29, 1: 132: 47, 2: 23: 61	8.10E-05
rs10944478	0: 370: 33.5, 1: 24: 65.5, 2: NA: NA	0.0879
rs1862893	0: 150: 46.5, 1: 191: 35, 2: 53: 23	0.0731
rs1111267	0: 206: 42, 1: 164: 28, 2: 24: 26	0.0126
rs4457655	0: 99: 55, 1: 199: 32, 2: 96: 33	0.2941
rs7157967	0: 314: 32, 1: 79: 55, 2: 1: 93	0.0441
rs11575845	0: 366: 38, 1: 28: 12.5, 2: NA: NA	0.0003
rs4963212	0: 261: 32, 1: 115: 38, 2: 18: 72.5	0.0443
rs6479789	0: 266: 29, 1: 104: 49, 2: 24: 50.5	0.0099
rs2023075	0: 294: 43, 1: 92: 23.5, 2: 8: 13.5	0.0091
rs28366162	0: 361: 38, 1: 33: 13, 2: NA: NA	0.0020

Although bootstrapping was hypothesised to be quicker than running CV, it actually took longer to run. This was due to the CV method consisting of 10 runs of 5 folds on the 40 blocks = 2000 PLS models, the bootstrapping consisted of 100 runs of the 40 blocks = 4000 PLS models. Therefore, until exploration using independent data can be performed the CV methodology in 6.4.2 will be continued to be used.

6.4.4. Using COREXpress to run SPLS or CCR

When using mixOmics to perform PLS modelling it requires substantial data pre-processing before the SPLS models can be fitted; such as splitting the GWAS into blocks of SNPs less than 12000 and imputing missing SNP data. To see if the modelling process could be more efficient, the software COREXpress (Magidson, 2011) was investigated. COREXpress is capable of running SPLS regression models on much larger blocks of data (reported to be up to 30,000 or more) which enables each chromosome to be fitted in a complete block. It also performs all data imputation as part of the SPLS modelling even when performing CV.

COREXpress can perform both SPLS regression and a new similar modelling approach called correlated component regression (CCR) which is soon to be included as a chapter in Abdi et al. (2013). Brief conclusions of this work are summarised below.

- Similar to mixOmics, SNPs with >20% missing data and SNPs with very low variation had to be removed prior to the analysis otherwise the program stopped working.
- COREXpress could model entire chromosomes at once (maximum attempted was 24374 variables).
- The methods used by COREXpress are similar to the 'percentage fold' method, whereby the number of times a variable is selected in each fold and each run, after specifying a minimum number of variables to keep in the model, is used. As shown in section 5.3.2, Table 5.2, this approach (particularly on the lower level models where there are few clear signals) is highly dependent on the number of variables you ask it to output. It was shown that if no minimum is selected it only output disease duration, if a minimum of X is selected then it output X variables each time making the final model dependant on an arbitrary cut off. The 'average rank' method as described in 6.4.1 creates models without having to specify these arbitrary limits.
- Although it is much faster to run a PLS model in COREXpress (5 minutes per chromosome block) it has no programming functionality and everything has to be done manually. Therefore considerable time is spent loading data and setting the models to run on each block of chromosomes and then outputting the results manually.

The results using SPLS 10 runs of 5-fold CV in mixOmics (section 6.4.2), SPLS in COREXpress and CCR in COREXpress were compared in Table 6.4. When the top 100 variables were examined for each model, only 14 of the SNPs were selected by all three models (rs1076933, rs10898438, rs1449760, rs1449779, rs1705013, rs1754068, rs2101201, rs470215, rs4770403, rs4898652, rs6814036, rs7319076, rs760609 and rs9847705). 44 SNPs were selected in 2/3 models and 148 SNPs each in only one model. Hence, all three models are selecting quite different SNPs.

Table 6.4 Comparison of mixOmics 10 run 5-fold CV model and COORExpress models

Final 100 variable model	mixOmics SPLS	SPLS COORExpress	CCR COORExpress
Number of SNPs/Env[a]	95 SNP, 5 Env	89 SNP, 11 Env	94 SNP, 6 Env
R^2	0.942	0.906	0.927
R^2-CV	0.882	0.871	0.899
% significant[b]	71.6%	87.6%	85.1%

[a] Number of SNPs and environmental variables retained in the top 100 for the final model
[b] % of SNPs which are statistically significant at the 5% level per section 4.2.5.

External replication could calculate what proportion of the SNPs selected by each model could be replicated and hence help to indicate which is the better model. However, an independent set of data is not available. It was therefore decided to take forward the methods from section 6.4.2 because of the flexibility to order the variable by an average rank, as opposed to specifying the number of variables to extract. In addition, it was felt much easier to be able to automate the running of the entire model process rather than manually having to import data, select the model fitting options individually for each chromosome and export the results.

6.5. 'Average rank' method results on 'GWAS SNPs' dataset

Using 10 runs of the 'average rank' method (as agreed in section 6.4.1 and 6.4.2) the top 505 ranked variables resulted in an R^2=0.942, R^2-CV =0.882 and correlation between the actual and predicted Larsen score was 0.953.

Unlike the 'percentage fold' method where 505 variables were selected because this was the number extracted in the top set of variables in 5/5 folds in all 50 runs, the 'average rank' method does not have a way of selecting the number of variables for the final model. It simply ranks the entire set of variables and 505 was used above simply for comparison with the 'percentage fold' method. Therefore, after developing the 'average rank' method which was a quicker method than the more standardly used 'percentage fold' method for fitting models to a large number of SNPs, it was now required to develop a method of selecting the most appropriate number of variables for the final model.

6.5.1. Number of variables required

It is highly unlikely that each SNP is contributing the same amount to the prediction model and as more SNPs are added, they should be explaining less and less variation in accordance with the way the SNPs were selected by the median rank of their loading. Therefore, how many variables are needed in the model to still form a good prediction? If fewer than 505 variables still predict well, then this would reduce the measures required to be collected for using the prediction model in the clinic.

To explore how the model prediction changes with a different number of variables, SPLS models were fitted to the 'GWAS SNPs' dataset, the coefficients of the model estimated and then used to predict the same set of patients. This was performed using the five environmental variables only, followed by the top 20, 50 and 100 variables, ranked according to the 'average rank' method. The correlation between actual Larsen score and predicted Larsen score increased from 0.605, 0.72, 0.812, up to 0.89 with 100 variables (Figure 6.11). As the model with 505 variables achieved a correlation of 0.953, the correlation is only improved by 0.063 by adding an extra 405 variables.

Therefore it appears that 100 (or perhaps even less) are sufficient for a good prediction model. The additional 405 are likely to be explaining random noise. Further methods are investigated in section 8.3 of how to select the optimum number of variables when using the 'average rank' method.

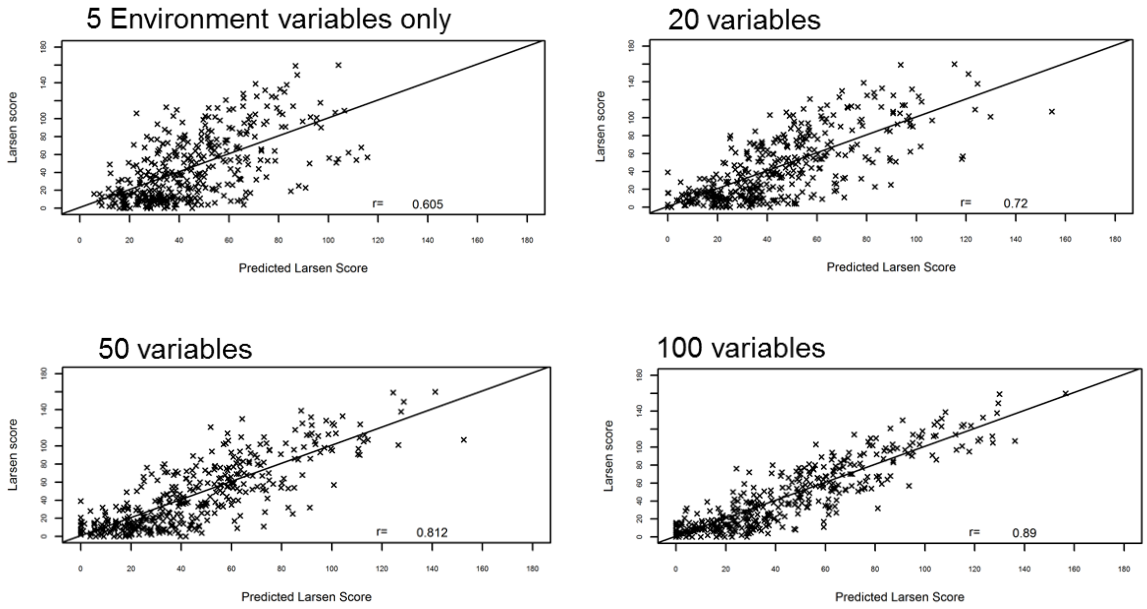


Figure 6.11 Relationship between number of variables in the model and correlation.

6.5.2. Interpretation of top 100 variables

The top 100 variables using the 'average rank' method were examined to assess if they had previously been identified in the literature reviews (section 2.3 and 2.4) as potential predictors of RA severity. Only five environmental variables were selected in the top 100 variables. Disease duration and symptom duration (selected as the 1st and 2nd most important variables) are highly correlated and clearly a large contributor to RA severity. The 3rd and 4th variables selected were age at time of diagnosis and age at onset of symptoms. These variables are also highly correlated and identified in the literature review (section 2.4.2.2). BMI was selected as the 9th most important predictor of RA severity and identified in the literature review in section 2.4.2.3.

Interestingly ACPA value did not appear in the top 100 variables and was actually ranked 208th. Although this is surprising given the general consensus that it defines two subgroups of RA disease (section 6.2.2), it may have a lack of sensitivity as ACPA was not recorded at disease diagnosis and therefore could have varied over time.

Based on the literature review (section 2.4) it is also surprising to find smoking duration and smoking pack years to be ranked 233rd and 439th respectively. Alcohol quantity consumed was also ranked very low at 645th.

It was also expected to find the HLA-DRB1 variables as coded by Tezenas du Montcel et al. (2005) S1, S2, S3d, S3p and X to be important as HLA-DRB1 was seen to be the most replicated genetic variant influencing RA severity (section 2.2.3). However, none of these variables were taken forward from the lower level by chromosome block models. To investigate possible reasons why they were not selected, the median Larsen score and p-value from a NB model (as described in

4.2.5) was extracted for DRB1 categories and none were found to be statistically significant when analysed univariately (Table 6.5). In fact, producing Table 6.5 on the 912 subjects analysed in Chapter 4, the ZINB LR p-values are also not significant at the 5% level (DRB1-S1: p=0.062, DRB1-S2: p=0.983, DRB1-S3d: p=0.886, DRB1-S3p : p=0.833 and DRB1-X : p=0.161). In conclusion, it appears this cohort of patients would not be able to identify HLA-DRB1 as a key predictor using univariate or multivariate modelling.

Table 6.5 Median Larsen score value for each DRB1 category (394 subjects)

Variable	N subjects	Median Larsen score	p-value	Observation
DRB1-S1 0	286	38.0	0.592	22.3% decrease from 0 to 1, 39.5% decrease from 0 to 2.
DRB1-S1 1	102	29.5		
DRB1-S1 2	6	23.0		
DRB1-S2 0	175	33.0	0.542	17.2% increase from 0 to 1, 39.1% increase from 0 to 2.
DRB1-S2 1	187	37.0		
DRB1-S2 2	32	44.5		
DRB1-S3d 0	371	36	0.321	No linear pattern
DRB1-S3d 1	21	55		
DRB1-S3d 2	2	14		
DRB1-S3p 0	181	35.0	0.823	No linear pattern
DRB1-S3p 1	189	36.0		
DRB1-S3p 2	24	36.5		
DRB1-X 0	257	38.0	0.661	No linear pattern
DRB1-X 1	115	32.0		
DRB1-X 2	22	35.5		

The details of each SNP in the final model are shown in Table 6.6. The table columns show the selection order (from one to 100, excluding the environmental variables), the number of times the SNP was selected out of 50 models (10 runs of 5-fold CV), the average rank (median order from the folds and runs), the number of patients, the median Larsen score for each genotype and p-value from a NB model as described in section 4.2.5.

Sixty-eight of the top 95 SNPs (71.6%) were significant at the 5% level and 43 (45.3%) at the 1% level. Investigation of the medians for each genotype in Table 6.6 revealed that most SNPs have evidence of a monotonic increase or decrease in median Larsen score. However, this summary statistic does not inform us of the underlying distribution of the data.

The following SNPs, identified using PLS methods to be in the top 100 variables of importance, were also found in genes documented in the literature review (section 2.3); rs470747 is an intron variant in the MMP1 gene and WTAPP1, rs470215 is an intron variant in WTAPP1 and 3' UTR variant in MMP1 gene, rs470168 is a 3' UTR variant in the MMP10 gene and rs1843933 is intron variant in the CARD11 gene. Only rs1076933 (Intron variant in ARHGAP8 and PRR5-ARHGAP8) selected in the chapter 4 modelling was also selected in this GWAS modelling. Attempts to externally validate selected important SNPs and further quantify whether the model is performing better than chance alone is investigated in sections 6.5.4 and chapter 8.

Table 6.6 Table of the top 100 selected SNPs from the 'average rank' method

SNP	Order/ n times selected/ avg rank	Genotype: N subjects : Median Larsen score	p-value	Chr.	Description of SNP
rs11211654	5/ 50/ 8	0: 343: 32, 1: 48: 58.5, 2: 3: 107	0.00123	2	A/G Intron variant in the TPO gene (thyroid peroxidase). (key enzyme involved in the thyroid hormone synthesis)
rs9326161	6/ 50/ 15	0: 350: 32.5, 1: 42: 64.5, 2: 2: 79.5	0.00082	2	C/T Intron variant in the TPO gene (thyroid peroxidase). (key enzyme involved in the thyroid hormone synthesis)
rs760609	7/ 50/ 18.5	0: 120: 25, 1: 211: 36, 2: 63: 58	0.00267	6	G/T intergenic variant of unknown function
rs1449779	8/ 50/ 19	0: 262: 29, 1: 106: 46.5, 2: 26: 63	0.00005	4	C/T intergenic variant of unknown function
rs2218495	10/ 50/ 27.5	0: 285: 29, 1: 101: 51, 2: 8: 49.5	0.01850	2	A/G intergenic variant of unknown function
rs7661111	11/ 50/ 29.5	0: 317: 41, 1: 74: 20.5, 2: 3: 2	0.00044	4	A/G Intron variant in the protocadherin 7 (PCDH7) gene.
rs9884240	12/ 50/ 36.5	0: 292: 33, 1: 98: 51.5, 2: 4: 85.5	0.01299	4	C/T intergenic variant of unknown function
rs754043	13/ 12/ 37	0: 364: 33, 1: 27: 58, 2: 3: 112	0.11438	16	C/T intron variant in the GPT2 gene. Glutamic pyruvate transaminase (alanine aminotransferase) 2.
rs4898652	14/ 50/ 37.5	0: 135: 25, 1: 189: 40, 2: 70: 49	0.00792	14	A/G intron variant in the SOS2 gene. Son of sevenless homolog 2 (Drosophila).
rs470747	15/ 50/ 38.5	0: 166: 27, 1: 182: 37.5, 2: 46: 60.5	0.05962	11	C/T intron variant of the MMP1 gene and WTAPP1.
rs470215	16/ 50/ 43	0: 166: 26.5, 1: 183: 38, 2: 45: 60	0.06504	11	A/G intron variant of WTAPP1. 3' UTR variant in MMP1 gene.
rs11897977	17/ 50/ 44.5	0: 332: 33, 1: 58: 54.5, 2: 4: 108	0.00975	2	A/G intron variant in TPO.
rs7142677	18/ 49/ 49	0: 330: 32, 1: 61: 59, 2: 3: 138	0.04593	14	A/C intron variant in MOK protein kinase gene (a member of the mitogen-activated protein kinase superfamily).
rs1558910	19/ 50/ 52	0: 258: 29, 1: 121: 49, 2: 15: 55	0.02358	3	C/T intron variant in the DGKG gene (diacylglycerol kinase, gamma 90kDa)
rs2015829	20/ 50/ 60	0: 262: 29, 1: 112: 50, 2: 20: 61	0.00007	4	C/T intergenic variant of unknown function
rs10031745	21/ 50/ 75	0: 293: 33, 1: 98: 42.5, 2: 3: 86	0.02259	4	C/T intron variant in PDE6B (phosphodiesterase 6B, cGMP-specific, rod, beta)
rs6814036	22/ 50/ 77	0: 128: 21, 1: 187: 39, 2: 79: 49	0.00138	4	C/T intergenic variant of unknown function
rs9563202	23/ 50/ 78.5	0: 317: 32, 1: 73: 56, 2: 4: 92	0.06573	13	G/T intergenic variant of unknown function
rs4609584	24/ 50/ 88.5	0: 336: 33, 1: 55: 58, 2: 3: 54	0.00238	11	A/G variant 7400 bases upstream of TEAD1 (TEA domain family member 1) and 6497 bases upstream of DD413619.
rs2933639	25/ 50/ 90.5	0: 132: 50.5, 1: 198: 33, 2: 64: 21	0.00175	5	C/T intergenic variant of unknown function.

SNP	Order/ n times selected/ avg rank	Genotype: N subjects : Median Larsen score	p-value	Chr.	Description of SNP
rs879606	26/ 50/ 96	0: 274: 41.5, 1: 112: 22, 2: 8: 27	0.15814	17	A/G variant 1328 bases upstream of PPP1R1B (protein phosphatase 1, regulatory (inhibitor) subunit 1B)
rs11299	27/ 50/ 98	0: 323: 40, 1: 69: 18, 2: 2: 9.5	0.00542	20	C/T 5' UTR variant of CSRP2BP (CSRP2 binding protein) and 3' UTR variant of PET117 (homolog S. cerevisiae)
rs1449760	28/ 50/ 98.5	0: 106: 23.5, 1: 199: 35, 2: 89: 54	0.01095	4	A/G intergenic variant of unknown function
rs1374153	29/ 49/ 103	0: 352: 33, 1: 42: 63, 2: NA: NA	0.00966	2	A/G intergenic variant of unknown function
rs955038	30/ 50/ 104.5	0: 277: 29, 1: 109: 56, 2: 8: 56	0.00526	2	A/G intergenic variant of unknown function
rs7948160	31/ 50/ 111	0: 244: 29, 1: 128: 49.5, 2: 22: 60	0.01112	11	A/G intron variant in SPON1 (spondin 1, extracellular matrix protein)
rs1754068	32/ 50/ 113.5	0: 215: 29, 1: 147: 38, 2: 32: 60.5	0.01200	9	A/G intergenic variant of unknown function
rs6788054	33/ 49/ 118	0: 299: 33, 1: 91: 49, 2: 4: 83.5	0.05727	3	C/T intron variant in DGKG (diacylglycerol kinase, gamma 90kDa)
rs2686386	34/ 50/ 124	0: 269: 29, 1: 117: 50, 2: 8: 71.5	0.11752	12	C/T variant 8007 bases upstream of P2RX4 (purinergic receptor P2X, ligand-gated ion channel, 4)
rs1375775	35/ 50/ 128	0: 229: 27, 1: 139: 44, 2: 26: 53	0.00018	4	A/G intergenic variant of unknown function
rs10506802	36/ 4/ 130.5	0: 366: 38, 1: 28: 12.5, 2: NA: NA	0.00002	12	G/T intron variant in SYT1 (synaptotagmin I)
rs7968671	37/ 4/ 130.5	0: 366: 38, 1: 28: 12.5, 2: NA: NA	0.00002	12	C/T intron variant in SYT1 (synaptotagmin I)
rs3751143	38/ 49/ 135.5	0: 271: 27, 1: 112: 54.5, 2: 11: 53	0.12724	12	G/T missense variant in P2RX7 (purinergic receptor P2X, ligand-gated ion channel, 7)
rs346650	39/ 50/ 136.5	0: 122: 47, 1: 186: 38, 2: 86: 19	0.00007	5	A/G intergenic variant of unknown function
rs1250043	40/ 49/ 137.5	0: 219: 27, 1: 151: 48, 2: 24: 51.5	0.01923	2	A/G intron variant in LINC00607 (long intergenic non-protein coding RNA 607)
rs1076933	41/ 49/ 142	0: 146: 48, 1: 187: 36, 2: 61: 18	0.01126	22	A/G intron variant in ARHGAP8 and PRR5-ARHGAP8
rs1323056	42/ 50/ 145.5	0: 143: 29, 1: 199: 36, 2: 52: 54.5	0.05706	6	A/G intron variant in HS3ST5 and BC042098
rs470168	43/ 49/ 146	0: 187: 29, 1: 169: 38, 2: 38: 62.5	0.22409	11	A/G 3' UTR variant in MMP10 (matrix metallopeptidase 10 (stromelysin 2))
rs12211935	44/ 49/ 147	0: 269: 28, 1: 114: 42, 2: 11: 67	0.00867	6	A/G intergenic variant of unknown function
rs1013940	45/ 50/ 150	0: 325: 32, 1: 66: 54, 2: 3: 99	0.00130	2	C/T 5' UTR missense variant in SLC5A7 (solute carrier family 5 (choline transporter), member 7)
rs1705013	46/ 50/ 155	0: 161: 47, 1: 193: 31, 2: 40: 26	0.03301	10	C/T intergenic variant of unknown function

SNP	Order/ n times selected/ avg rank	Genotype: N subjects : Median Larsen score	p-value	Chr.	Description of SNP
rs10878920	47/ 49/ 157.5	0: 286: 32, 1: 97: 51, 2: 11: 55	0.01212	12	A/G intergenic variant of unknown function
rs1843933	48/ 49/ 166	0: 201: 42, 1: 165: 33, 2: 28: 15.5	0.01497	7	A/G intron variant in CARD11 (mRNA-caspase recruitment domain family, member 11)
rs10898438	49/ 50/ 168	0: 121: 28, 1: 191: 33, 2: 82: 58	0.01379	11	A/G intergenic variant of unknown function
rs9847705	50/ 49/ 170	0: 206: 45, 1: 156: 27.5, 2: 32: 22	0.01438	3	G/T intergenic variant of unknown function
rs799923	51/ 50/ 171	0: 228: 43.5, 1: 145: 27, 2: 21: 17	0.06218	17	A/G intron variant in BRCA1, associated with early onset breast cancer.
rs3970313	52/ 47/ 174	0: 343: 33, 1: 51: 54, 2: NA: NA	0.00551	4	A/G intergenic variant of unknown function
rs2132046	53/ 50/ 176	0: 289: 40, 1: 99: 22, 2: 6: 10	0.02793	7	A/G intergenic variant of unknown function
rs1532190	54/ 50/ 190	0: 118: 54, 1: 184: 33.5, 2: 92: 23	0.31945	3	G/T intron variant in GXYLT2 (glucoside xylosyltransferase 2)
rs2017317	55/ 49/ 193.5	0: 143: 49, 1: 189: 36, 2: 62: 18	0.01530	22	C/T intron variant ARHGAP8 and PRR5-ARHGAP8
rs7770227	56/ 50/ 194.5	0: 134: 28.5, 1: 206: 37, 2: 54: 52.5	0.06430	6	A/G intron variant in HS3ST5 and BC042098
rs10498671	57/ 50/ 195.5	0: 267: 44, 1: 111: 29, 2: 16: 21	0.00890	6	C/T intron variant in BMP6 (bone morphogenetic protein 6)
rs354082	58/ 14/ 214	0: 364: 33, 1: 30: 58, 2: NA: NA	0.07122	7	C/T variant 1363 bases downstream of TRNA-Cys
rs4770403	59/ 48/ 216.5	0: 244: 30, 1: 133: 44, 2: 17: 66	0.00708	13	A/G 5’ UTR variant in SGCG (sarcoglycan, gamma 35kDa dystrophin-associated glycoprotein)
rs8015527	60/ 45/ 220	0: 332: 33, 1: 60: 51.5, 2: 2: 112	0.00315	14	C/T intergenic variant of unknown function
rs892533	61/ 49/ 221	0: 194: 41.5, 1: 168: 31, 2: 32: 16	0.00021	13	A/G intergenic variant of unknown function
rs6427859	62/ 49/ 225.5	0: 252: 30.5, 1: 132: 43.5, 2: 10: 45	0.05745	1	A/C intron variant in CAMSAP2 (calmodulin regulated spectrin-associated protein family, member 2)
rs7528217	63/ 49/ 225.5	0: 252: 30.5, 1: 132: 43.5, 2: 10: 45	0.05745	1	C/T intron variant in CAMSAP2 (calmodulin regulated spectrin-associated protein family, member 2)
rs265005	64/ 48/ 229	0: 198: 27.5, 1: 169: 40, 2: 27: 68	0.00402	5	C/T intergenic variant of unknown function
rs2646852	65/ 50/ 230.5	0: 134: 24.5, 1: 188: 37.5, 2: 72: 55	0.10964	1	A/G intergenic variant of unknown function
rs8032023	66/ 49/ 234	0: 126: 24, 1: 195: 39, 2: 73: 54	0.00135	15	C/T intron variant in RORA (RAR-related orphan receptor A)
rs1354969	67/ 48/ 234	0: 131: 54, 1: 198: 33, 2: 65: 25	0.13326	8	C/T variant 3454 bases upstream of U6.

SNP	Order/ n times selected/ avg rank	Genotype: N subjects : Median Larsen score	p-value	Chr.	Description of SNP
rs6050732	68/ 42/ 235	0: 359: 33, 1: 34: 64, 2: 1: 118	0.01084	20	A/G intron variant in ZNF337
rs10511820	69/ 49/ 236.5	0: 184: 44, 1: 160: 33, 2: 50: 22	0.05292	9	A/C intron variant in LINGO2 (leucine rich repeat and Ig domain containing 2)
rs2889450	70/ 48/ 238	0: 146: 49, 1: 186: 36.5, 2: 62: 19	0.04118	2	C/T intergenic variant of unknown function
rs11902332	71/ 47/ 239	0: 294: 33, 1: 88: 39, 2: 12: 100.5	0.90722	2	A/G intron variant in ANKRD44 (Ankyrin repeat domain 44)
rs2242310	72/ 49/ 241	0: 122: 51.5, 1: 192: 37, 2: 80: 21	0.00604	4	C/T intron variant in SCFD2 (sec1 family domain containing 2). 3' UTR variant of AK055055.
rs4623925	73/ 49/ 244	0: 220: 44, 1: 146: 26.5, 2: 28: 17	0.00168	11	A/G intron variant in BC070093.
rs508835	74/ 49/ 244	0: 220: 44, 1: 146: 26.5, 2: 28: 17	0.00168	11	C/T intron variant in BC070093.
rs2101201	75/ 49/ 244.5	0: 119: 52, 1: 191: 37, 2: 84: 21	0.00744	4	C/T intron variant in SCFD2 (sec1 family domain containing 2)
rs10457277	76/ 49/ 246	0: 268: 28.5, 1: 116: 41.5, 2: 10: 64	0.00776	6	A/G intergenic variant of unknown function
rs3115667	77/ 49/ 248.5	0: 244: 42.5, 1: 131: 33, 2: 19: 13	0.00039	6	A/G variant 9760 bases upstream of GPANK1, 5556 bases downstream of CSNK2B, 3172 bases downstream of LY6G5B and 1062 bases downstream of LY6G5C
rs782933	78/ 48/ 251.5	0: 320: 32, 1: 72: 55, 2: 2: 99.5	0.00121	15	A/C intron variant in RORA (RAR-related orphan receptor A)
rs7816057	79/ 49/ 253.5	0: 185: 26, 1: 175: 47, 2: 34: 54.5	0.03880	8	G/T variant 3498 bases upstream of U6.
rs2499937	80/ 49/ 255.5	0: 226: 44, 1: 139: 25, 2: 29: 19	0.00041	11	A/G intergenic variant of unknown function
rs4099560	81/ 48/ 256	0: 344: 33, 1: 48: 57.5, 2: 2: 83.5	0.47210	7	A/G intron variant in MKLN1 (muskelin 1, intracellular mediator containing kelch motifs)
rs2429862	82/ 47/ 256.5	0: 222: 28.5, 1: 151: 44, 2: 21: 73	0.00487	11	A/G variant 4628 bases downstream of BC040894 and 2862 bases upstream of RPLP0P2.
rs7852101	83/ 1/ 258	0: 367: 35, 1: 27: 73, 2: NA: NA	0.42565	9	A/C intergenic variant of unknown function
rs2843157	84/ 48/ 258.5	0: 308: 32, 1: 81: 59, 2: 5: 20	0.00236	1	A/G 3' UTR variant in SKI (v-ski sarcoma viral oncogene homolog (avian))
rs2680182	85/ 49/ 259	0: 336: 32, 1: 54: 57, 2: 4: 68	0.01280	2	A/G intron variant in LINC00607 (long intergenic non-protein coding RNA 607)
rs6916278	86/ 36/ 268.25	0: 361: 38, 1: 33: 13, 2: NA: NA	0.00143	6	A/G intron variant in LY6G6F (lymphocyte antigen 6 complex, locus G6F) and ABHD16A
rs7111383	87/ 49/ 271.5	0: 227: 43, 1: 139: 26, 2: 28: 17	0.00184	11	C/T intron variant in BC070093
rs7581781	88/ 35/ 272	0: 360: 33, 1: 34: 66, 2: NA: NA	0.01281	2	A/G intergenic variant of unknown function

SNP	Order/ n times selected/ avg rank	Genotype: N subjects : Median Larsen score	p-value	Chr.	Description of SNP
rs12001157	89/ 47/ 275.5	0: 349: 33, 1: 45: 59, 2: NA: NA	0.00502	9	A/G intron variant in APBA1 (amyloid beta (A4) precursor protein-binding, family A, member 1)
rs10792830	90/ 49/ 280.5	0: 131: 51, 1: 183: 34, 2: 80: 25	0.02843	11	A/G intergenic variant of unknown function
rs1860404	91/ 49/ 287.5	0: 291: 30, 1: 98: 52.5, 2: 5: 66	0.00043	10	C/T intron variant in SLC18A2 (solute carrier family 18 (vesicular monoamine), member 2)
rs2395672	92/ 46/ 291	0: 253: 29, 1: 126: 54.5, 2: 15: 44	0.02697	6	A/G intron variant in FTSJD2 (FtsJ methyltransferase domain containing 2)
rs1413882	93/ 48/ 298.5	0: 245: 32, 1: 135: 39, 2: 14: 79	0.14185	13	A/G intergenic variant of unknown function
rs6420308	94/ 48/ 298.5	0: 245: 32, 1: 135: 39, 2: 14: 79	0.14185	13	C/T intergenic variant of unknown function
rs10925085	95/ 50/ 299.5	0: 149: 46, 1: 172: 34, 2: 73: 24	0.31166	1	C/T missense variant in OR2G2 (olfactory receptor, family 2, subfamily G, member 2)
rs2643891	96/ 48/ 300	0: 275: 33, 1: 110: 51, 2: 9: 22	0.06654	1	A/G intron variant in MORN1 (MORN repeat containing 1)
rs1879390	97/ 48/ 302.5	0: 309: 33, 1: 81: 55, 2: 4: 54.5	0.00380	12	G/T intergenic variant of unknown function
rs702479	98/ 47/ 303	0: 192: 28, 1: 178: 39, 2: 24: 67.5	0.05599	7	C/T intergenic variant of unknown function
rs7319076	99/ 48/ 303.5	0: 110: 22, 1: 190: 37, 2: 94: 54	0.03406	13	A/G intergenic variant of unknown function
rs2272593	100/ 49/ 312.5	0: 245: 42, 1: 132: 33, 2: 17: 13	0.00073	6	A/G missense variant in PRRC2A (proline-rich coiled-coil 2A)

6.5.3. Variance partitioning of the final model

The top 100 variables of the model resulted in a correlation of r=0.890 between actual and predicted Larsen score. However, how much of this model's predictive ability was due to genetics and how much was due to disease duration, symptom duration and the environmental variables?

Using the multi-block variance partitioning method (Skov et al., 2008) described in section 4.3.4, the variation that the model explained was partitioned into that attributable to disease duration and symptom duration, environment and genetics. 56% of the explained variation by the model was attributable to genetics alone using the 'GWAS SNPs' dataset to produce a final model containing 100 SNPs (Table 6.7, Figure 6.12). This is a considerable increase from the 'all subjects' dataset model which investigated 368 SNPs in chapter 4 and found approximately 13% of the explained variation was attributable to genetics alone.

Table 6.7 Multi-block variance partitioning of 100 variable model using 'average rank' method

First block fitted	A: Larsen score variation explained by each block when fitted alone (%)	B: Larsen score variation unique to that block i.e. the variation is not common to the other blocks (%). (Percentage unique: B/A*100).
DD&SD	32.64	17.98 (55.1%)
Environment	26.94	15.77 (58.5%)
Genetics	75.51	56.23 (74.5%)

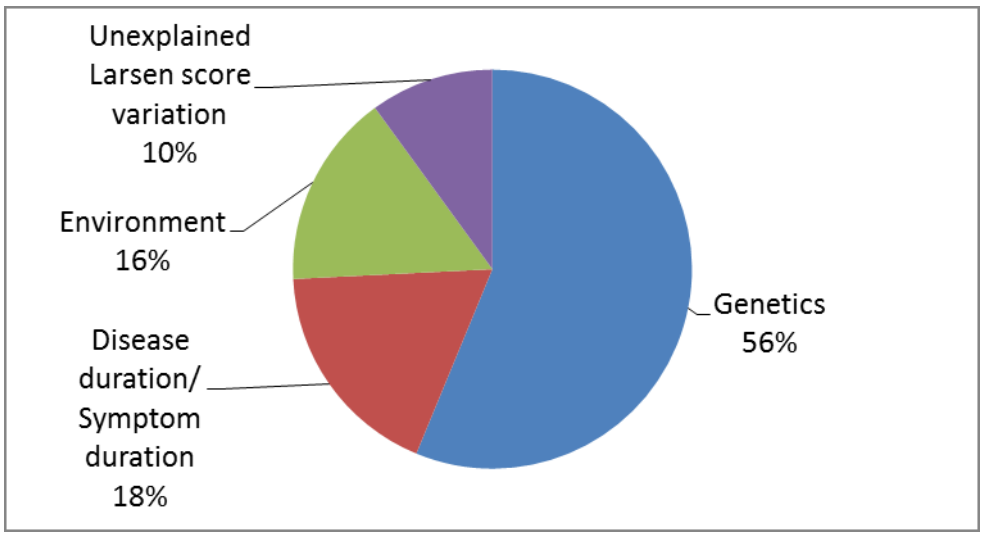


Figure 6.12 Pie chart of variance partitioning of 100 variable model using 'average rank' method

6.5.4. External replication of the top 10 SNPs

In order to determine whether the top 10 SNPs identified as being predictive using the 'average rank' method on the 'GWAS SNPs' dataset (Table 6.6) are transferable to other cohorts, independent replication was provided by Dr Annette van der Helm-van Mil and Hanna van Steenbergen of the Leiden University Medical Center (LUMC). Data consisted of a GWAS performed on 290 ACPA-negative patients from the Leiden Early Arthritis Clinic (LEAC) (van Aken et al., 2003) and a GWAS performed on 385 ACPA positive patients from the North American Rheumatoid Arthritis Consortium (NARAC) (Croiseau and Cordell, 2009). As these were longitudinal data in an early RA cohort, log transferred Sharp scores were analysed using a repeated measures analysis of variance (ANOVA) including a genotype by time interaction term in addition to other potential confounders (such as gender, age, ACPA, RF, smoking and BMI). The choice of covariates was in accordance with LUMC standard practices for analysing these cohorts. The Sharp score is similar to the Larsen score except that it includes a count of joint space narrowing in addition to erosive damage. In the case where GoRA SNPs did not exist in the NARAC or LEAC cohorts, proxy SNPs were used where possible (Table 6.8).

Table 6.8 Replication results of the top 10 SNPs (proxy SNPs identified)

SNP	NARAC results	LEAC results
rs11211654	NA	NA
rs9326161	NA	NA
rs760609	rs1323056: p=0.34	rs1323056: p=0.12
rs1449779	p=0.49	NA
rs2218495	p=0.95	NA
rs7661111	p=0.71	p=0.69
rs9884240	rs9997286: p=0.30	p=0.17
rs754043	NA	NA
rs4898652	rs7141809: p=0.11	p=0.10
rs470747	NA	p=0.08

NA = Not available

Three of the top 10 SNPs were not available even as proxies in either of the cohorts. Of the remaining seven SNPs, only rs4898652 with p-values in NARAC and LEAC of 0.11 and 0.10 respectively and rs470747 with a p-value of 0.08 in the LEAC cohort looked promising. The NARAC and LEAC data for rs4898652 and the LEAC data for rs470747 indicated a decrease in severity scores with an increase in genotype (0 to 1 to 2). However, the GoRA 'GWAS SNPs' dataset, indicated an increase in severity scores with an increase in genotype (Figure 6.13). It was therefore concluded that none of the top 10 SNPs could be externally replicated at this time. This could be due to the model being formed on a relatively small number of patients (N=394), or that the replication sets are in early RA and do not have similar characteristics to the GoRA population. Unfortunately external replication using a PLS model cannot be performed due to no cohort existing with the same SNPs as used in the GoRA dataset. Chapter 8 provides further exploration into methods of model validation.

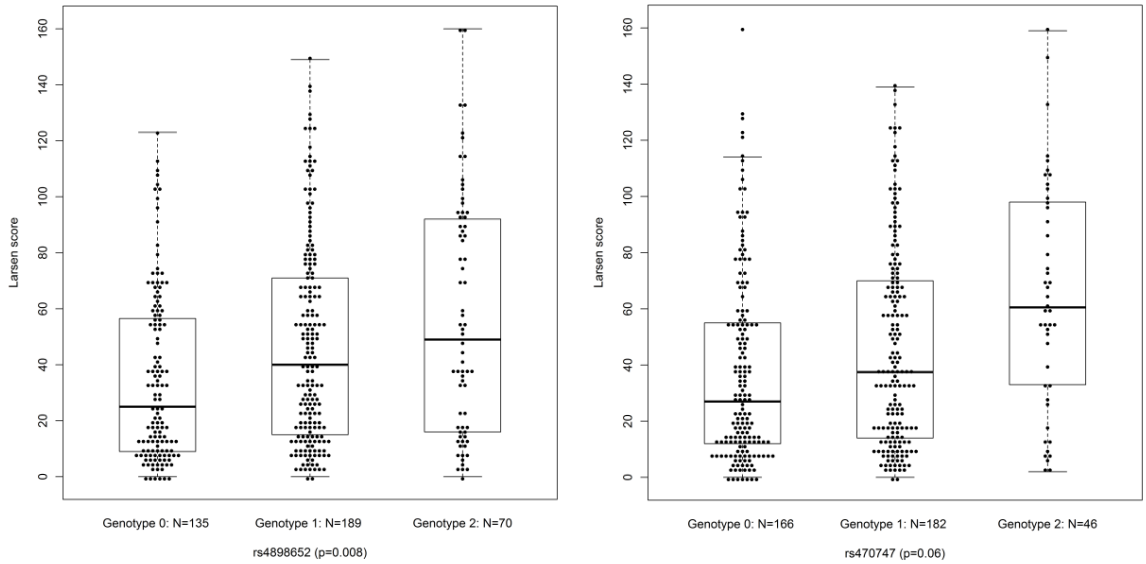


Figure 6.13 Raw data, box plots and p-values from GoRA study for rs4898652 and rs470747

6.5.5. Manhattan plot using NB models

When analysing the top 100 SNPs selected by the SPLS methodology, none of the SNPs in Table 6.6 met univariate genome wide significance of $p<1.5x10^{-7}$, calculated using a bonferroni correction for 325000 SNPs. Sixty-eight of the top 95 SNPs (71.6%) selected were significant at the 5% level and 43 (45.3%) were significant at the 1% level using NB models as described in 4.2.5.

Manhattan plots are used in the literature to identify the strongest associations between phenotypes and SNPs. It was decided to investigate whether the SPLS 'average rank' method was missing any highly statistically significant SNPs which would have been identified by a more standard univariate analysis presented in a Manhattan plot.

SNPs were fitted one at a time in a NB model (as described in section 4.2.5). The p-values from univariately testing all 324563 SNPs were plotted using a Manhattan plot (Figure 6.14) using code sourced from: http://people.virginia.edu/~sdt5z/0STABLE/qqman.r

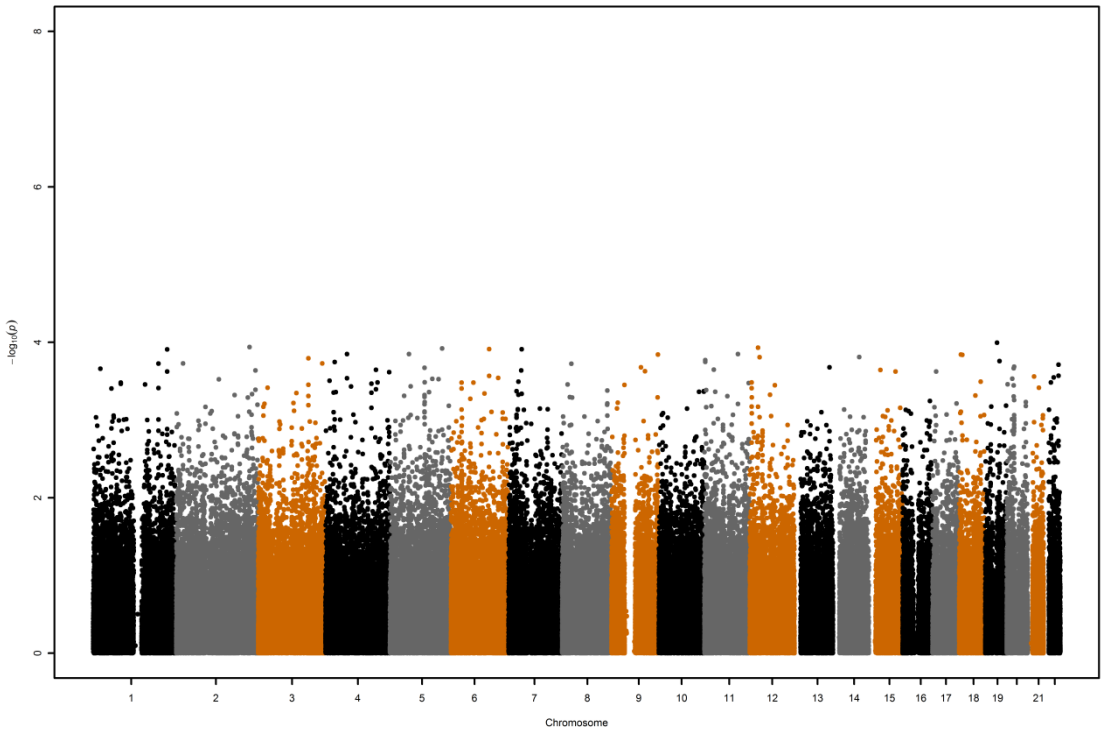


Figure 6.14 Manhattan plot of 'GWAS SNPs' dataset for 324563 SNPS.

In Figure 6.14 none of the p-values from the NB models are less than $1.5x10^{-7}$ (Above 7 on the y-axis). Given the testing of 324563 SNPs using a sample size of 394, it is of interest to calculate the power for the effect sizes observed in this data. The power is the probability of rejecting the null hypothesis, given the null hypothesis is false. In other words rejecting that $\beta=0$ given the true value for $\beta\neq0$.

It was decided to use the Poisson distribution to provide an over-simplified approximation of the power which could be achieved using ZINB models. For ease of calculation, it was assumed the models only have the one variable of interest (one SNP) in the model. The equations shown below

describe how PASS 2008 sample size calculation software estimates power for Poisson regression models (Hintze, 2008) using formulae derived by Signorini (1991).

Using the Poisson distribution, the probability of y events can be modelled as follows, where u represents the mean incidence rate:

$$\Pr(Y = y|\mu) = \frac{e^{-\mu}\mu^{y}}{y!}$$

In Poisson regression, it is assumed that μ is determined by a set of regressor variables. For estimation of a SNPs effect on the Larsen score, $\mu = \exp(\beta_0 + \beta_1 x_1)$ where x_1 can be thought of as the SNP of interest (a single covariate in the model), β_0 is the baseline response rate and exp(β_1) is the change in Larsen score for a one point increase in frequency of the common allele (i.e. from a genotype of 0 to 1). In the regression model, we want to test the null hypothesis that β_1=0 (no SNP effect) versus the alternative that β_1=B1. Hintze (2008) use the following formula to calculate the sample size (N):

$$N = \phi \frac{\left(Z_{1-\alpha/2}\sqrt{V(b_1|\beta_1 = 0)} + Z_{1-\beta}\sqrt{V(b_1|\beta_1 = B1)}\right)^2}{e^{\beta_0} B1^2}$$

In the above equation, α is the type I error, β is the type II error, Z is the standard normal deviate and ϕ is a measure of over-dispersion, where for these calculations, no over-dispersion is assumed by setting ϕ = 1. As $\beta_1$is unknown, b_1 is estimated from the data using the maximum likelihood estimate of β_1. $V(b_1|\beta_1 = 0)$ is the variance of b_1 under the null hypotheses and $V(b_1|\beta_1 = B1)$ is the variance of b_1 under the alternative hypothesis. In order to calculate the variance of $b_1$under the null and alternative hypotheses, a probability distribution for the SNP (x_1) being investigated must be specified. The binomial distribution was selected. By converting the genotypes to allele frequencies, the proportion of patients (p_{x1}) with the common allele can be calculated. PASS software uses this value in the calculations of the variance of $b_1$shown below.

Under the Null Hypothesis: $V(b_1|\beta_1 = 0) = \frac{1}{p_{x1}(1-p_{x1})}$

Under the Alternative Hypothesis: $V(b_1|\beta_1 = B1) = \frac{1}{1-p_{x1}} + \frac{1}{p_{x1}e^{B1}}$

The SNPs analysed in the Manhattan plot (Figure 6.14), were used to calculate the frequency of common allele. On average, across all 324563 SNPs, the common allele was observed in 75% of patients (p_{x1}=0.75). However, SNPs vary widely in the common allele frequency, hence p_{x1}=0.5 was also investigated as this would estimate the power using the most optimistic assumption about p_{x1}.

Using the SNPs analysed in the Manhattan plot (Figure 6.14), the maximum effect size associated with a change in the frequency of the common allele (from a genotype value of 0 to 1) was observed to be a 145% increase in the Larsen score. However, this would be an exceptionally large effect size for one polymorphism. A range of effect sizes are presented in Figure 6.15 and Figure 6.16 to correspond to a 10% increase (0.1) up to a 145% increase (2.45). The calculations provided are likely to be optimistic as over-dispersion ($\phi \neq 1$) was not allowed for in the power calculation. The power estimate would be reduced if over-dispersion was present.

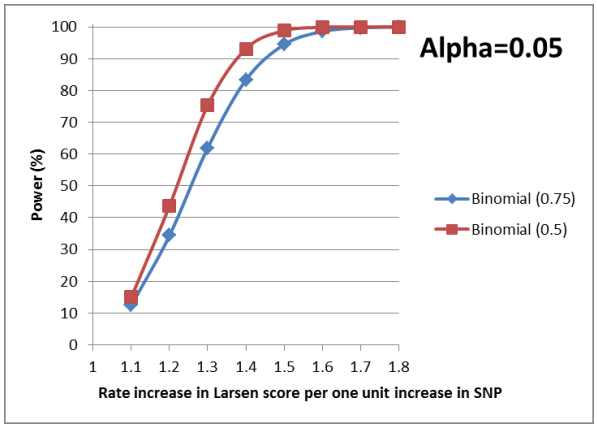


Figure 6.15 Power calculations (alpha=0.05)

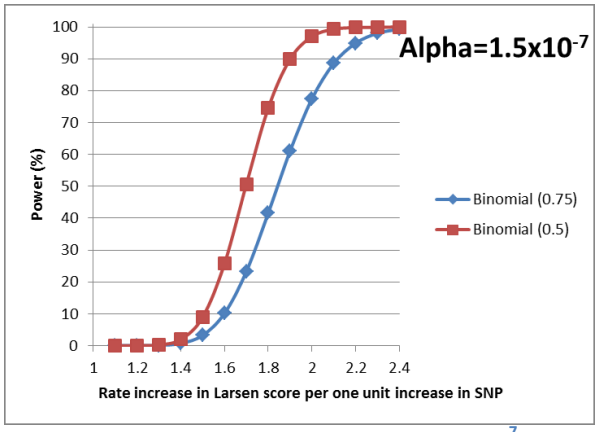


Figure 6.16 Power calculations (alpha=1.5x10^{-7})

Assuming the testing of a single SNP using 394 subjects, a 5% alpha level of significance corresponds to more than 90% power to detect a 50% increase (1.5) in the Larsen score for a one point common allele frequency increase (0 to 1) (Figure 6.15). However, when testing 324563 SNPs, the GWAS alpha significance level was <1.5x10^{-7}. In this scenario, less than a 10% power is associated with a 50% increase for a one point common allele frequency increase (0 to 1) (Figure 6.16). Therefore, with 324563 SNPs and 394 subjects, we require substantially bigger SNP effect sizes (potentially unrealistically large for the effect of one SNP) to have enough power to detect significance at the GWAS adjusted alpha level. It is worth noting that this GWAS SNP data formed part of a much larger GWAS study consolidating many subjects together from different cohorts.

The top 95 SNPs with the lowest p-values from the NB models were extracted and compared to the 95 SNPs selected in the final 10 run 5-fold CV model (section 6.5.2). The following six SNPs were chosen by both models, however, the remaining 89 were different: rs1375775, rs1860404, rs2499937, rs3115667, rs7661111 and rs892533. The reason for the difference could be due to PLS identifying SNPs correlated with severity, aiming to explain all the variation as opposed to a univariate analysis which identifies SNPs with the largest effect sizes.

It was decided to use the smallest p-values from the NB models (combined with the top environmental variables chosen using SPLS), to see how well such a model can predict the Larsen score. The expectation was that selecting SNPs based on the smallest p-values from NB models would perform worse than using SPLS and the 'average rank' method. The reason for this is that SPLS chooses variables to explain as much variation as possible, whereas univariate p-values are simply identifying the SNPs with highest association to the Larsen score.

To investigate how the 95 SNPs with the lowest p-values in the NB models perform in a PLS prediction model, they were used along with the same five environmental variables identified in the 'average rank' method SPLS model to predict the Larsen score. Figure 6.17 reveals a correlation of 0.845 which only very slightly lower to the correlation observed in 'average rank' method with 100 variables, r=0.890 (Figure 6.11). Examination of the two correlation plots indicates that the univariate SNP selection model has a higher residual error associated with the higher Larsen scores (Figure 6.17). In comparison, the SPLS model has a constant residual error (Figure 6.11). This is thought to be because variables selected through PLS models will attempt to explain all Larsen

score variation using many predictors in combination, whereas univariate testing uses the variables in isolation. Whilst, the SPLS 'average rank' method appears to be the better approach, this research identifies the difficulty selecting true predictors from such high-dimensional data. Further validation methods are therefore investigated for the 'GWAS SNPs' dataset in section 8.3.

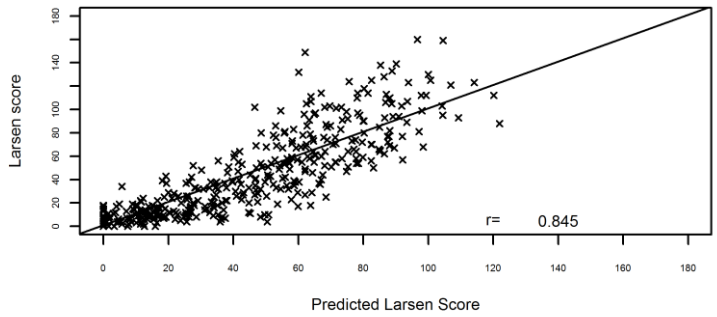


Figure 6.17 Actual versus predicted Larsen score for 95 SNPs selected from Manhattan plot

Using the p-values from the NB models to produce a Q-Q plot indicates that observed p-values are higher (larger) than would be expected for this number of tests (Figure 6.18). For example the lowest p-value we observe is approximately 0.0001 ($1x10^{-4}$) whereas with this number of tests we would expect 0.000001 ($1x10^{-6}$). The figure was recreated excluding SNPs with low MAF (<5%, <10% and <20%), however all figures looked similar to Figure 6.18. Therefore, it was unlikely that the observed departure from the line of the theoretical distribution is due to low MAF. As the line begins to depart from the theoretical distribution when the observed p-values are just 0.1 (1 on the Y-axis), it is speculated that the assumptions for the NB distribution do not hold well for this data. In section 4.3.2 and 4.4.2, p-values calculated using ZINB models on the 'All subjects' data were found to vary widely just by changing the method of data imputation which resulted in sparse categories having present or absent data. Further issues with using standard regression theory with multivariate data is further discussed in 3 and 10.3.

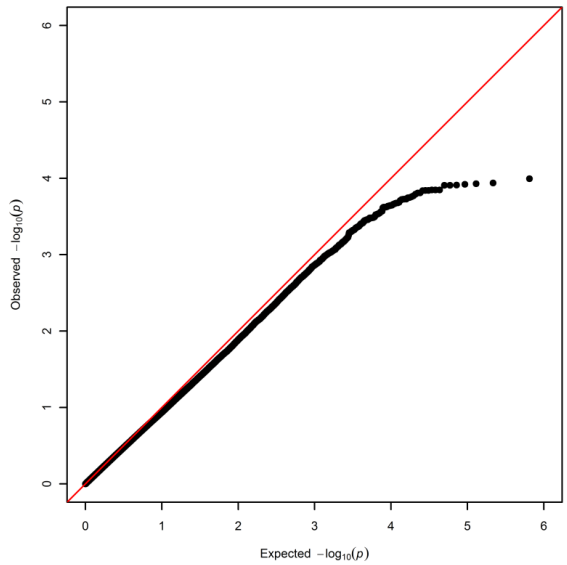


Figure 6.18 Q-Q plot of NB models for 'GWAS SNPs' dataset (324563 SNPS)

6.6. Summary

The SPLS 'percentage fold' method which was developed in chapters 4 and 5 was applied to the 'GWAS SNPs' dataset. Due to the size of the dataset (325482 SNPs), SNPs were divided into 40 blocks, with each chromosome being split so that no block had more than 12000 SNPs in it. Each block of SNPs were modelled including the environmental variables each time. The 'percentage fold' method consisted of taking each block in turn and assessing it for the optimum number of variables to extract, by plotting the R^2-CV by number of variables extracted. Numbers of variables to extract for each chromosome block ranged from 200 variables to just 10. 50 runs of 5-fold CV, extracting the relevant number of variables each time, were produced on all 40 blocks. Any variables being selected in at least 2 out of 5 folds in all 50 runs were retained for the higher level model (maximum SNP model section 6.3.2.2). As so many SNPs were carried through (919), it was also decided to produce a minimum SNP model (section 6.3.2.1) which only retained SNPs for the higher level model if they were extracted in 5 out of 5 folds in all 50 runs. It was intended that the CV process would be repeated on the higher level models to reduce the number of variables further, however, investigation under CV suggested in both cases, keeping all variables formed the best model. The minimum SNP model (493 SNPs) resulted in a correlation between actual and predicted Larsen score of r=0.950 compared to the maximum SNP model (N=919) which was only slightly better r=0.963. This suggested the additional 426 SNPs were probably over fitting the model.

Various efficiencies to the model fitting strategy were examined in order to reduce the time needed to run the models and remove the manual step of reviewing plots in order to choose the number of variables to extract (section 6.4). As an alternative to the time consuming 'percentage fold' method, which was being used in SPLS literature often after pre-filtering of SNPs (Le Cao et al., 2011), the 'average rank' method was developed. This method used the absolute value of each variable's loading and calculated a median rank over the CV folds and runs of the data. This removed the requirement to extract a set number of variables and allowed a more sensitive ranking strategy for the SNPs. SNPs which were rare but important could feature high in some models but not others due to the fold of the patients. The new strategy meant that these SNPs now had a chance of being retained for the higher level model. Once variables were ranked, the top 200 were selected from every chromosome block and taken through to the higher level model. The same process was then used to rank the variables in the higher level model and final models using the top 5, 20, 50 and 100 variables were examined for their predictive ability (section 6.5.1). It appeared that whilst adding variables up to 100 substantially increase the predictive ability and resulted in a correlation between actual and predicted Larsen score of r=0.890, the additional gain in prediction from 100 to 505 variables was probably due to over fitting of the model and only achieved a correlation of r=0.953 (increase of 0.063).

Testing revealed that instead of performing 50 runs of 5-fold CV, a similar model was produced using just 10 runs. This along with using the 'average rank' method reduced the SPLS running time from approximately a week (with manual intervention) to eight hours.

In conclusion, the 'GWAS SNPs' dataset can be successfully modelled in an automated way using SPLS in mixOmics in under eight hours on a standard computer. Unfortunately, there was considerable data preparation work required to split the SNPs into blocks and impute missing data. COREXpress software was investigated, however as it used the 'percentage fold' method, the models created were highly dependent on the number of variables you asked it to retain in the model. In addition, COREXpress has no programming language. Fitting each of the chromosomes individually and having to manually select the model to run was very time consuming. A further way to reduce the preparation work and running time is investigated in chapter 7 using SIMCA software.

Variance partitioning of the final 100 variable model suggested 56% of the Larsen score variation was explained by the 95 SNPs in the model. 18% was explained by disease duration and symptom duration,16% was explained by age at time of diagnosis, age at onset of symptoms and BMI, leaving just 10% unexplained. Clearly these estimates are very high and may suggest an over fitted model. 68 of the 95 SNPs (71.6%) were significant at the 5% alpha level and 43 (45.3%) at the 1% level when analysed using NB univariate models. Median Larsen scores by genotype tended to reveal monotonic increases or decreases indicating an observable correlation between Larsen score and the top 100 SNPs. However when the top 10 SNPs were selected and analysed in an independent cohort using a univariate analysis, none of the findings could be replicated. Unfortunately, external validation using SPLS methods are not possible as there are no available independent cohorts with the same SNPs recorded. Chapter 8 examines other internal methods of validation to attempt to quantify the ability of the model to predict the Larsen score.

It was decided to test all of the SNPs, one at a time using NB models and present the results in a Manhattan plot. No SNPs met the genome wide significance level, perhaps due to the sample size being too small (N=394). SNPs with the 95 lowest univariate p-values were entered into a PLS model, which contained the same top five environmental variables, to investigate whether this alternative way of selecting SNPs can achieve similar or better prediction than SPLS. A correlation between actual and predicted Larsen score of r=0.845 was obtained which was almost as good as the PLS prediction model (r=0.89) despite only six SNPs overlapping both models.

The final SPLS model achieved excellent predictions of the Larsen score (correlation of 0.89 with 100 variables). However, how the model performs on an independent dataset is questionable. It is standard practice to use CV to reduce the chance of over fitting of the model and estimate how well the model is likely to fit on an independent set. However, the final model was selected from 10 runs of 5-fold CV and as there were so many SNPs to select from, the variables which were chosen (in worst case), could be variables which describe the Larsen score very well under CV in this sample. Therefore prediction estimates should be treated with extreme caution. Quantification of the over fitting is investigated in Chapter 8.

7. SIMCA modelling of Larsen score – 'GWAS SNPs' dataset

7.1. Aims

The aim of this chapter is to:

- Investigate the added functionality and any efficiencies of fitting SPLS in SIMCA compared to mixOmics to determine the best method to use in future.
- Investigate orthogonal PLS (OPLS) and compare it to PLS in SIMCA and mixOmics.

7.2. Methods

A limitation of the analysis methods developed in section 6.4.2 is that mixOmics required the 'GWAS SNPs' dataset to be split into smaller blocks due to reaching R's internal memory limit. This resulted in variable selection methods being applied during the model fitting so that only the most predictive variables were carried forward to the higher level model (section 6.4.1). To do the variable selection, CV was required and this needed imputation of missing values to be completed prior to the analysis. This splitting of data into blocks may result in the loss of information for correlated variables which happen to appear in different blocks. In addition, the prior imputation of missing data may result in false precision of modelled data as imputed missing values are treated as real values in the analysis. Although CORExpress (section 6.4.4) could model larger blocks and no prior imputation was required, it still required a lot of data manipulation and the variables selected for the final model were very different to mixOmics perhaps due the inflexibility of the variable selection methods.

SIMCA by Umetrics AB Version 13.0 (Eriksson et al., 2006a, Eriksson et al., 2006b) is a multivariate data analysis software capable of analysing the entire 'GWAS SNPs' dataset as one dataset and requires no prior imputation. This saves considerable time in data preparation prior to analysis. The basic PLS methodology is the same as used by mixOmics however some key differences are summarised in Table 7.1.

As SIMCA analyses all the SNPs together in one model, methods derived in section 6 are not required (such as 'percentage fold' and 'average rank' methods). Unlike the mixOmics function which took approximately five hours to run SPLS on the 'GWAS SNPs' dataset on a multi cluster machine or eight hours on a standard computer, SIMCA can run PLS models on the entire 'GWAS SNPs' dataset in approximately 1.5 hours. The reduction in time using SIMCA is likely to be due to it running one PLS on all of the 325,482 SNPs together which is considerably faster than mixOmics which takes 40 blocks of SNPs, fits 10 runs of 5-fold CV, selects the most important variables to carry forward to the higher level model and then fits the final model.

As all SNPs are analysed together, SIMCA performs simultaneous PLS & NIPALS imputation of missing data. During the calculation of the PLS components, SIMCA sets the residuals for the missing values to zero. The missing data is estimated iteratively using the minimum distance projections onto the current estimate of the loading and score vector. This has the effect of missing data having no influence on the model. Whilst this appears to be a more appropriate method of handling missing data, care is required to not simply analyse variables which are unreliable due to the quantity of missing data (Pedreschi et al., 2008). For this reason, it is useful in SIMCA to use the

'missing value tolerance' option which removes observations or variables which have a pre-defined amount missing data (e.g. greater than 20%).

Table 7.1 Comparison of R mixOmics and SIMCA® for PLS, SPLS and OPLS

	R mixOmics v3.1 and 4.0-2[a]	SIMCA® Umetrics AB Version 13.0[b]
PLS – Using no variable selection	✓	✓
SPLS – The most important predictive X variables are selected for final model and only these variables contribute to the components	The 'Percentage fold' method (Section 6.2.4) and 'Average rank' method (6.4.1) have been developed for use with mixOmics.	No variable selection but perhaps could use the size of the final loadings to select top X variables for final model.
OPLS – Uses no variable selection and the variation in X is split into that related to Y and that orthogonal to Y. The 1st component contains all X variation related to Y (See below).	Not currently available	If >1 component was required (say for multiple Y variables) then this would give improved interpretation as only the 1st component is necessary to view relationship between X and Y
Cross validation	Data split into folds, the left out rows of X and Y are predicted by the model	Data split into folds, the left out rows of X and Y are predicted by the model.
Imputation required?	Must impute missing data prior to modelling if using CV	Simultaneously imputes missing data and can perform PLS under CV
Maximum number of variables which can be modelled	<12000 at a time due to reaching R's internal memory capacity	Modelled all 325,482 variables at once and a maximum was not found

a: R Foundation for Statistical Computing v2.13.1 or v2.15.1 (Vienna, Austria)
b: (Eriksson et al., 2006a, Eriksson et al., 2006b)

In addition to having PLS functionality, SIMCA has implemented a recent extension to PLS, designed specifically to aid interpretation of prediction models when there is only one Y variable. Orthogonal PLS (OPLS) (Trygg and Wold, 2002), filters the variation in the X variables into that which is related linearly to the Y variable and that which is orthogonal (unrelated) to the Y variable. With a single Y variable, the first component of the model contains all of the information from X which is predictive of Y and hence if the model is required for prediction, only the 1st component is of interest. Subsequent components can therefore be ignored as they have no (or little in the presence of missing data) relation to Y. Fonville et al. (2010) recommend this approach as it increases the ability to model the variation of interest. The approach has also been recently used in other areas of prediction (Vajargah et al., 2012, Genneback et al., 2013). This is the only software known to perform OPLS although there are functions available in MATLAB and R which perform Kernel-OPLS for modelling non-linear relationships particularly for calibration studies (Bylesjo et al., 2008). However, as modelling of the Larsen score has only required one component to date, unless more components are required when analysing multiple RA severity variables, then this option is unlikely to be of much benefit.

SIMCA default options are used. These consist of NIPALS imputation for missing values, mean centring, unit variance for each variable and 7-fold CV whereby every 7th observation is assigned to the same fold. The CV is only used to estimate the fit of the final model. Although SIMCA has no variable selection criteria, to mirror the 'average rank' method used in mixOmics, the loadings vector from just one run using all subjects (not part of the CV), are used to quantify the influence of each X variable on the Larsen score. The 100 absolute largest loadings are extracted by exporting

the loading vector and sorting them in excel. These values are used to select the top 100 variables for the final model. This model is then compared to the mixOmics model which was described in section 6.5.2.

7.3. Results of SIMCA PLS and OPLS compared to mixOmics SPLS

The additional predictive ability of including more than one component in the model using the R^2 and R^2-CV was examined. For mixOmics PLS and SIMCA PLS only one component was deemed necessary. For SIMCA OPLS, all variation associated with the Y variable is contained in the first component therefore other components can be calculated but only the first component is ever needed to quantify a variables influence on the Y variable.

To enable comparison with the mixOmics SPLS model created using the 'average rank' method (section 6.5), the top 100 variables from SIMCA models using PLS and OPLS were compared. All three models retained the same five environmental variables of disease duration, symptom duration, age at onset of symptoms, age at time of diagnosis of disease and BMI. Table 7.2 reveals much overlap between the three models with80 / 95 SNPs being the same in all three models.

Table 7.2 Similarity of variables selected by mixOmics, SIMCA PLS and OPLS

Number of SNPs the same in each pair of models (out of 95)	mixOmics 10 run SPLS	SIMCA PLS
SIMCA PLS one run of 7-fold CV	83	--
SIMCA OPLS one run of 7-fold CV	81	89

Using one run in SIMCA of 7-fold CV created a very similar model to the mixOmics 40 blocks of 10 runs of 5-fold CV but took substantially less time to run (one hour compared to eight hours). It was decided to explore whether one run of 7-fold CV in mixOmics would result in a similar model. Unfortunately, there was a substantial loss in similarity. Although the five environmental variables selected were the same, the one run model had a maximum of 70 SNPs the same as the mixOmics 10 run model, the SIMCA PLS or the SIMCA OPLS. Only 64 SNPS were the same across all four models. Therefore, reducing the number of runs in the mixOmics modelling to just one run is likely to produce a poorer model. It is believed that SIMCA is more stable because it analyses all SNPs at the same time (imputing data using NIPALS on the full dataset, within the model), whereas mixOmics is split into smaller lower level blocks and the multiple runs of CV is used to determine which variables to carry forward to the higher level model.

In addition to SIMCA creating a very similar model to mixOmics in a much quicker time, it also had the added advantage of easy to produce graphics. By pasting in chromosome locations for each SNP into SIMCA, any graph can be automatically labelled by chromosome.

To demonstrate the graphical capabilities of SIMCA, a variable influence on projection (VIP) plot was produced for the SIMCA PLS model (Figure 7.1). The Y-axis of the figure indicates the size of influence each variable on the X-axis has on the Larsen score. In models with more than one component (or more than one Y variable), the advantage of using the VIP score is that the importance of the X variables are assessed using one value rather than multiple values for each component. VIP is derived using the weighted sum of squares of the PLS weights from each component contributing to the prediction of the Y variable(s). In general, variables with values above approximately 0.8 or 1 are considered to have an important influence on the Y prediction.

Figure 7.1 reveals that as expected the top four variables with highest VIP (all above 2) are disease duration, symptom duration, age at onset of symptoms and age at time of diagnosis. After these four variables, the remaining 95 SNPs in the model and BMI have a very similar contribution to the prediction of the Larsen score with a VIP score of between 0.8 and 1.2. SIMCA uses the seven models (from the 7-fold cross validation) to calculate standard errors and confidence intervals using the standard formula for jack-knifing as described by Efron and Gong (1983). 95% confidence limits are plotted around the estimated VIP score in Figure 7.1. These confidence limits could also be used to select variables, with preference given to those of large influence and narrower limits.

The direction of a variables contribution to the Larsen score can be investigated using either the loading for each variable (one per component extracted), or the model coefficients for each variable (calculated using all model components). Figure 7.2 shows the loadings for the 1st component from the SIMCA PLS model. It reveals disease duration and symptom duration to have a positive relationship with the Larsen score (a larger disease duration or symptom duration results in a larger Larsen score). At the right hand side of the x-axis, it reveals that as expected, age at onset of symptoms and age at time of diagnosis have a negative relationship with the Larsen score (a smaller age at onset of symptoms and time of diagnosis results in a larger Larsen score). BMI is revealed to also have a negative relationship based on the 'GWAS SNP' data suggesting a lower BMI results in a higher Larsen score. Each SNP can be interpreted in the same way. SNPs to the left of the plot, indicate that a one point increase in genotype, corresponds to an increase in Larsen score. SNPs to the right of the plot, indicate that a one point increase in genotype, corresponds to a decrease in Larsen score.

An alternative way to view variable loadings for each component is in a scatter plot. The first two components for the model using the SIMCA PLS (Figure 7.3) are displayed despite only the 1st component being required in the model. The second component is included for visual purposes only. To gauge the contribution to the Larsen score, a line is drawn from the 'Larsen_bas' variable (shown on the far right of the figure) through the origin. Each of the X variables are then projected onto this line (at a 90° angle) and the distance each projected variable is from the origin, indicates that variables contribution to the Larsen score. The further from the origin indicates a larger contribution. Figure 7.3 demonstrates that the top 95 SNPs all have very similar influence on the Larsen score. If all GWAS SNPs were plotted, they would have filled in the gap between the two groups of positive and negative highest influence SNPs covering the origin.

To demonstrate how the SIMCA OPLS may provide an advantage to PLS interpretation, a scatter plot of the loadings from the 1st component from the SIMCA OPLS model are shown in Figure 7.4. The vector of loadings corresponding to the first component is already rotated so that the 'Larsen_bas' (Larsen score) lies exactly on the X axis. The distance along the X-axis now corresponds to the size of contribution to the Larsen score and the Y axis can be ignored as this corresponds to variation orthogonal to the Larsen score. In this scenario, Figure 7.4 adds little extra visually than can be observed in Figure 7.3, however, if more than one component was significant in the model, having all variation associated with the Y axis projected onto one plane would be beneficial.

7.4. Summary

This chapter investigated whether SIMCA provided better functionality and similar results to the mixOmics macros previously used. SIMCA is substantially quicker than mixOmics and requires very little pre-processing of the data. Although the process cannot be fully automated as in R, it is relatively simple to import the 'GWAS SNPs' dataset and extract the top 100 variables using the absolute loading for each variable after sorting them in excel. The final 100 variables have to then be imported back into SIMCA to obtain the final model. The graphics are very easy to produce and can reveal patterns in the data particularly when colour coding is used for other identifiers (such as the chromosome identifier).

Both the PLS and OPLS models created in SIMCA using one run of 7-fold CV were similar to using 10 runs of 5-fold CV in mixOmics, with 80/95 variables overlapping all three models. However, a single run in mixOmics using 7-fold CV is not advised, as variable selection becomes more unstable with only 70/95 overlapping. The only advantage found to using OPLS compared to PLS is that the interpretation of the results would be simpler if there were multiple components required. The only drawback of using SIMCA compared to R is that it is very time consuming to select or deselect patients and variables (you have to individually click on each one). Hence, it is difficult to determine the optimum number of variables to include in a final model as no variable selection is available. In conclusion, mixOmics may be more desirable if you want to fit multiple datasets or split your patient sample into training/test sets, as once the pre-processing is completed, the whole process can be automated. However, SIMCA is better if you are modelling a single dataset as you do not have to perform any pre-processing or imputation of missing data and it has better graphics.

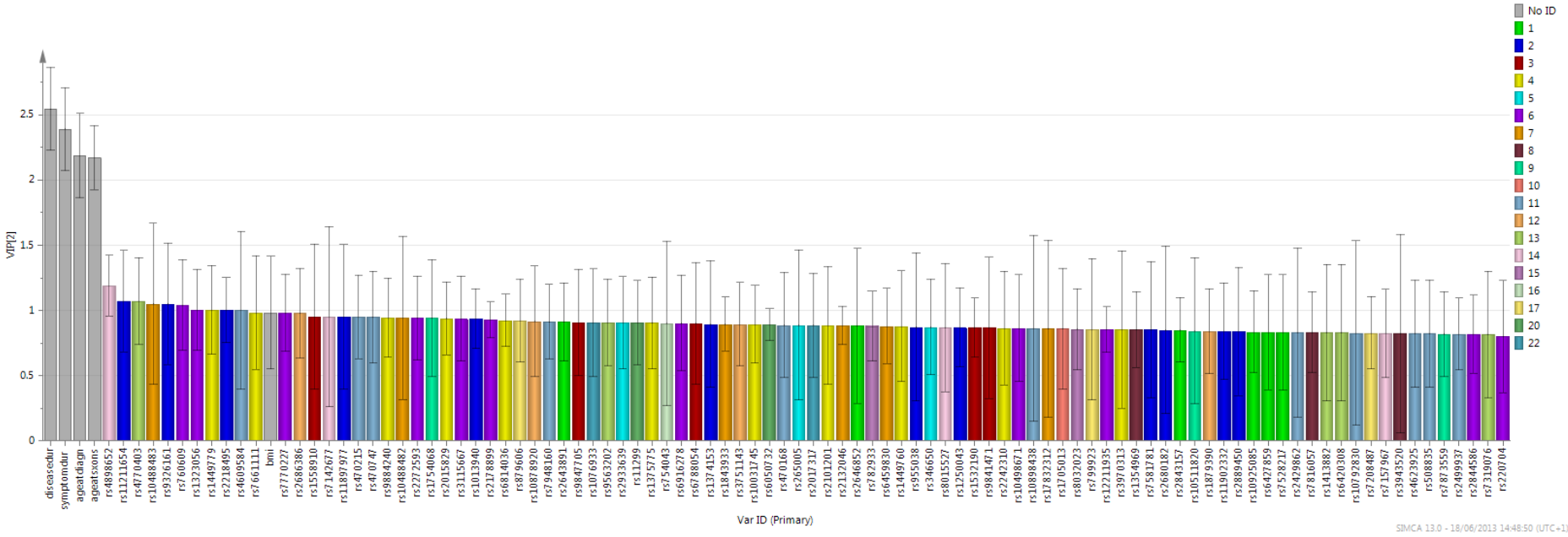


Figure 7.1 Variable importance plot for SIMCA PLS model with SNPs coded by chromosome.

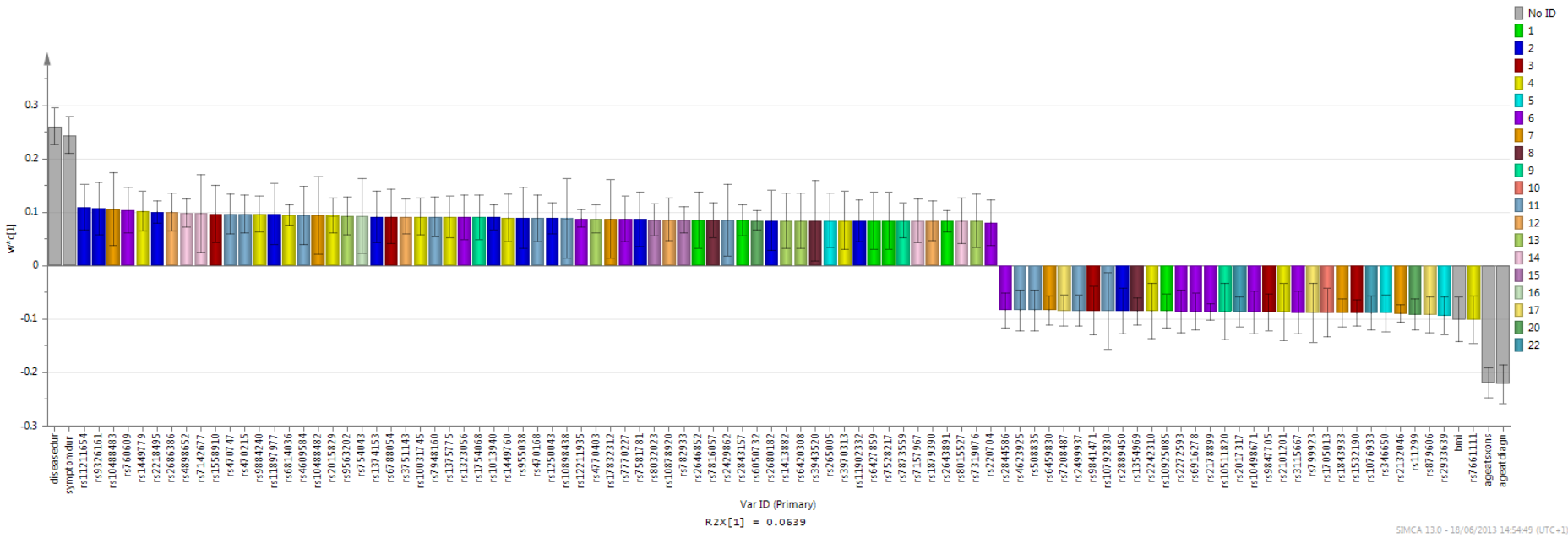


Figure 7.2 Loadings for SIMCA PLS model with SNPs coded by chromosome.

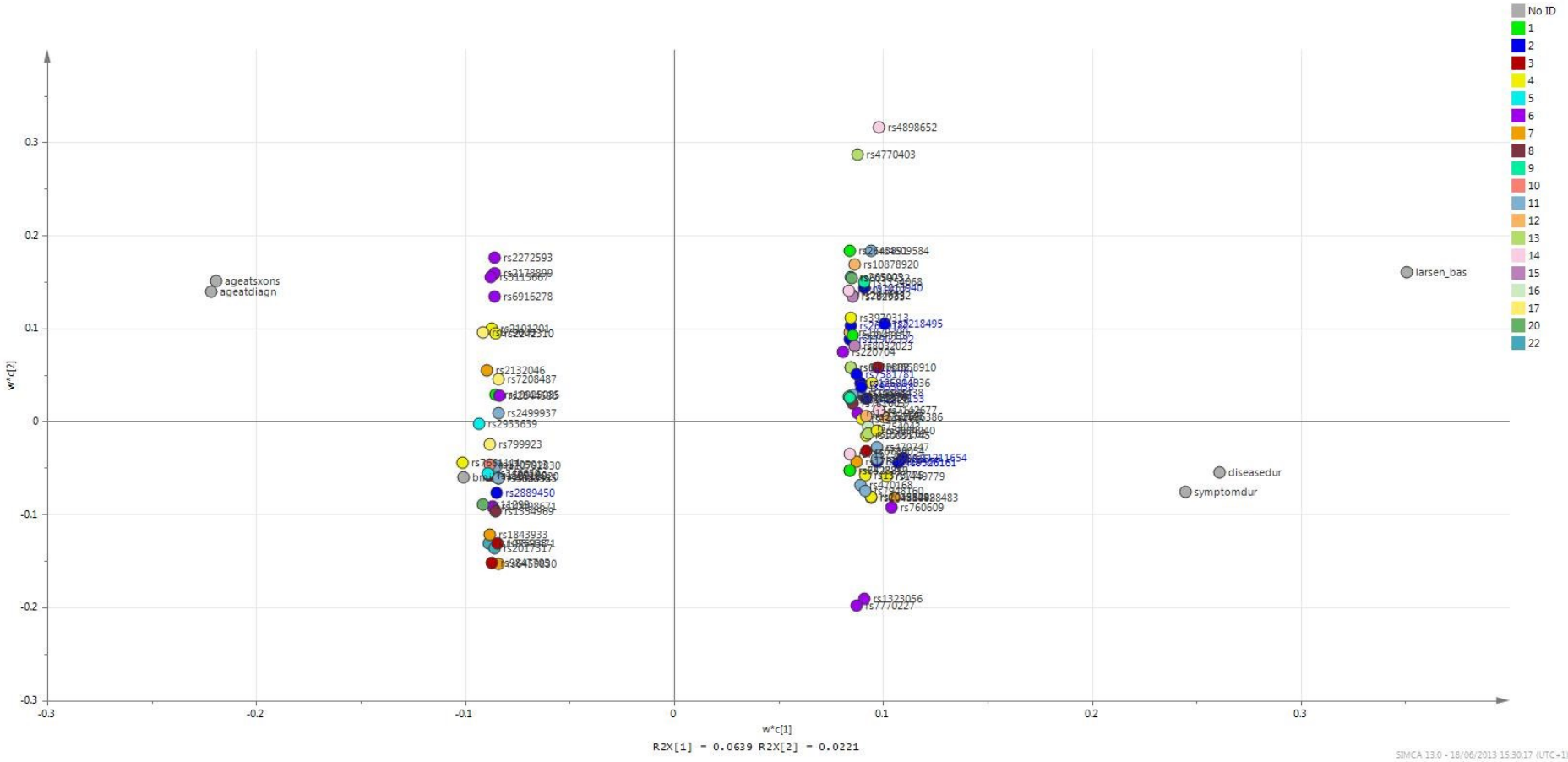


Figure 7.3 Scatter plot of loadings for SIMCA PLS model with SNPs coded by chromosome.

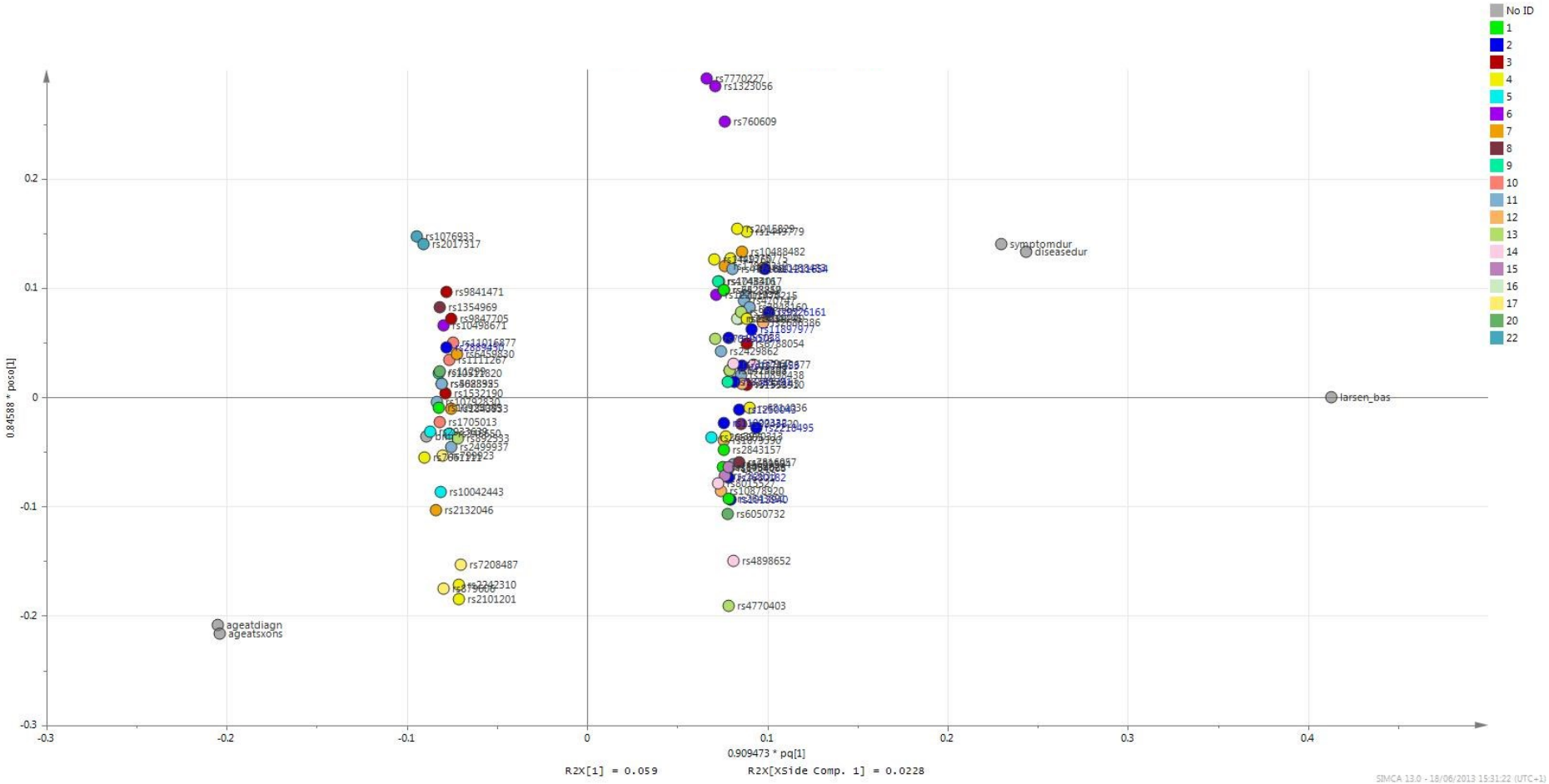


Figure 7.4 Scatter plot of loadings for SIMCA OPLS model with SNPs coded by chromosome.

8. SPLS regression of Larsen score: Further methods and validation

8.1. Aims

The aim of this chapter is to:

- Randomly permute the Larsen score data for the 'GWAS SNPs' dataset in order to investigate the performance of the model on unrelated Y and X data.
- Randomly permute the Larsen score and Environmental variables for the 'GWAS SNPs' dataset in order to investigate the additional contribution SNPs have on the predictive ability of the model.
- Split the 'GWAS SNPs' dataset into a training model dataset (80% of subjects) and a test only dataset (20% of subjects) and apply the 'average rank' method to estimate how the model may perform on an independent dataset. Investigate whether a 'two stage average rank' method could both order and select the optimum number of variables for the final model.
- Split the 'all subjects' dataset into a training model dataset (80% of subjects) and a test only dataset (20% of subjects) to estimate how the model may perform on an independent dataset. Investigate extending the 'average rank' method to a 'Two stage' or 'Three stage' process where stage 1) select the order of importance for variables, stage 2) select the optimum number of variables for the final model and stage 3) test the model on an independent set of patients. Investigate whether the model performs better if the 'all subjects' dataset is restricted to subgroups of disease, such as disease duration less than 10 years, less than 15 years or ACPA positive disease only.

8.2. Permutations Analysis – 'GWAS SNPs' dataset

8.2.1. Methods

Section 6.5 demonstrated a running time efficient, predictive model, using the 'GWAS SNPs' dataset to predict the Larsen score (the 'true' Larsen score data). However, it is important to determine whether the observed predictive ability of the model could have been achieved by chance, due to variable selection being performed from 325,482 SNPs and environmental variables, to predict just 394 subjects. Accurate prediction could have been obtained simply because of the high-dimensionality of the dataset. It was therefore decided to investigate the probability of achieving this predictive ability when there is no relationship between the X and Y variables.

The Larsen score data is randomly permuted 100 times whilst keeping all of the X variables linked to the original subject identifiers. SPLS is fitted to each of the 100 permuted datasets using the 'average rank' method. For each permutation, the top ranked 100 variables are used to calculate the R^2 and R^2-CV. 100 variables were chosen based on the findings in section 6.5.1. The distribution of the R^2 and R^2-CV values from the 100 models are compared to the true Larsen score data. The position of the R^2 and R^2-CV from the true Larsen score model is then used to calculate how likely it is to get this result, or more extreme, under the assumption of no relationship between the X and Y data.

A further permutation analysis is performed, randomly permuting the Larsen score and environment data together. This retains the relationship between the Larsen score and variables such as disease duration, symptom duration, ACPA, smoking and alcohol use however removes the

relationship with the genetic SNPs. Using 100 permutations, SPLS models are fitted using the 'average rank' method. As above, for each permutation, the top ranked 100 variables are used to calculate the R^2 and R^2-CV and the distribution of the R^2 and R^2-CV is compared to the true Larsen score data. If the permuted data performs as well as the real data, this indicates that the SNPs are not adding anything to the model above being selected by chance alone.

This work is completed using a Linux based high performance computing cluster 'Iceberg' and an updated mixOmics package (González et al., 2011, Lê Cao et al., 2009) version 4.0-2 as available on the 9th October 2012 in the R Foundation for Statistical Computing, Vienna, Austria (version 2.15.1).

8.2.2. Results

Using the 'average rank' method, the true Larsen score data was re-run through the entire modelling process. The top 100 variables were retained for the final model and resulted in an R^2 = 0.866 and an R^2-CV = 0.753.

The Larsen score data was randomly permuted 100 times and each new dataset consisting of the original X and permuted Y data was run through the entire modelling process calculating the R^2 and R^2-CV on the final model. Figure 8.1 demonstrates that on average there was no correlation (r=0) between the permuted and true Larsen score datasets, with a range of correlations between approximately ± 0.12. Figure 8.2 and Figure 8.3 show that when retaining 100 variables for the final models, none of the 100 permutations achieve an R^2 or R^2-CV greater than when the true Larsen score data is modelled. This suggests that generating a model as good or better than this one is unlikely ($p<0.01$), when the null hypothesis of no relationship between the X and Y variables is true. Therefore there is evidence that the true Larsen score data model contains variables predictive of the Larsen score.

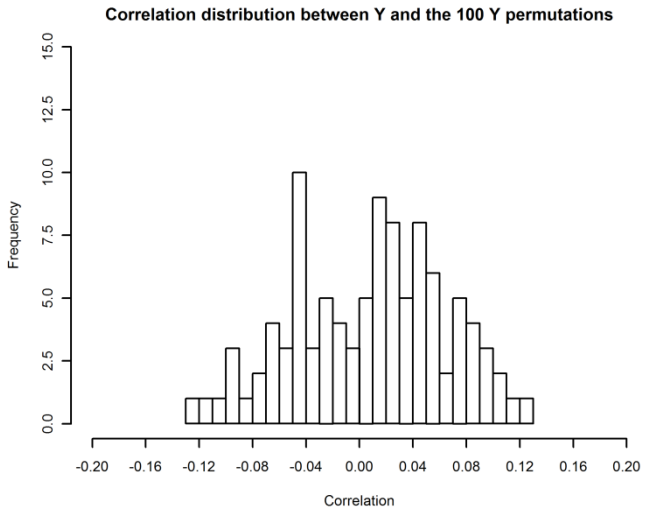


Figure 8.1 Correlation between 'real' and 100 permuted Larsen score datasets.

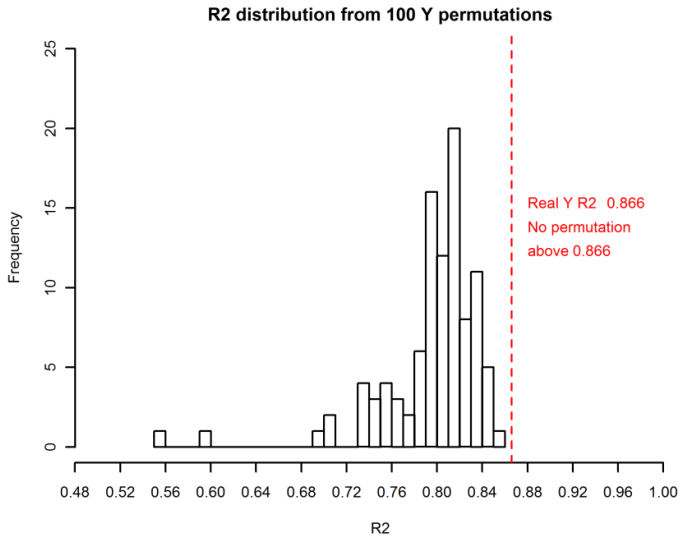


Figure 8.2 Distributions of R^2 for the 100 permuted Larsen score datasets

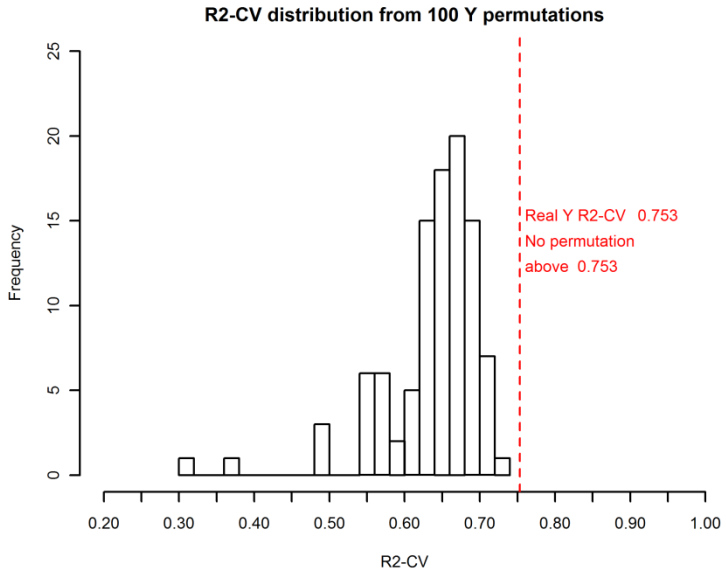


Figure 8.3 Distributions of R^2-CV (Q^2) for the 100 permuted Larsen score datasets

The Larsen score and environmental data were permuted 100 times and merged to the non-permuted SNP data. These new datasets were run through the entire modelling process using the top 100 selected variables as the final model. The R^2 and R^2-CV from each of the 100 final models were calculated. Figure 8.4 and Figure 8.5 reveal that two of the 100 permutations achieve an R^2 or R^2-CV greater than when the true Larsen score data is modelled. This suggests that modelling the Larsen score, using the environmental data alone, is unlikely to be able to achieve as good a predictive model as modelling the Larsen score using the environmental and genetic SNP data (p=0.02). Hence this analysis suggests that the SNP data is contributing to the model more than just by chance alone.

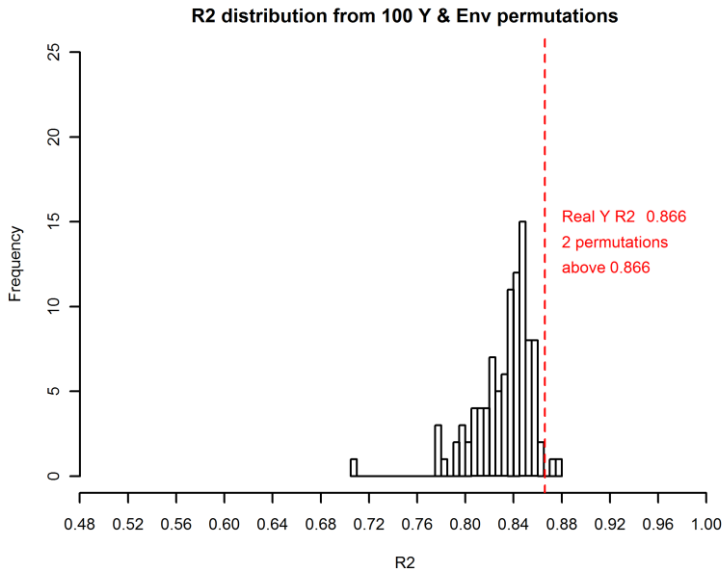


Figure 8.4 Distributions of R^2 for the 100 Larsen score and environment permuted datasets

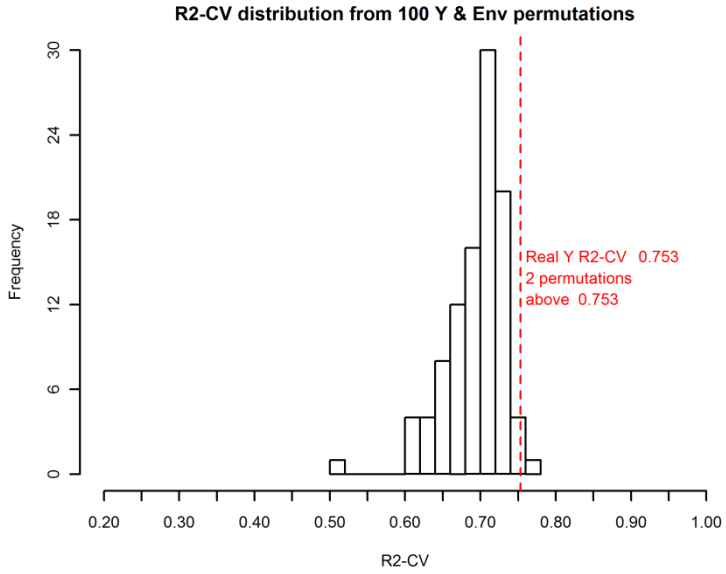


Figure 8.5 Distributions of R^2-CV (Q^2) for the 100 Larsen score and environment permuted datasets

8.3. Independent training and test sets – 'GWAS SNPs' dataset

8.3.1. 'Average rank' method using a separate training and test set

The 'average rank' method (described in section 6.4.1 and results shown in section 6.5) created an order of importance for each of the variables considered for the model. This order could then be used to form a final model which in section 6.5 contained the top 100 variables. It was initially anticipated, that the use of CV in the 'average rank' method, would prevent against over fitting as justified in 4.2.1.4. However, this could not be investigated, as all data was used to form the final model. The correlation between the actual and predicted Larsen score is highly likely to be an overestimation of model performance, because the same patients used to create the model have been predicted from the model.

A more reliable estimate would be obtained using a test sample which is not used at all in the creation of the model (Eriksson et al., 2006a, p. 377). It was decided in section 4.2.1 not to split the data into a test and training sample due to only having 394 subjects with GWAS data. However, after the CV and permutation tests resulted in very positive results, it is of interest to investigate how well an independent test set can be predicted. As no known other cohort of data exist with the same SNPs and Larsen score measurement, splitting the GoRA sample is the only option.

Eriksson et al. (2006a, p.377) recommend establishing a test set which represents the entire span of possible X data with good representation of the Y data as well. Although it is difficult with so many X variables to ensure good representation of the X's, subjects can be representative of the distribution of the Larsen score Y variable. Therefore, the subjects are sorted in ascending order by the Larsen score and every fifth subject, starting with the 3rd subject, is extracted for the test sample (20%). This test sample should then not be used at all in the creation of a model.

Using the 'average rank' method, the training set (80% of subjects) is used to form a model containing 100 variables (based on section 6.5.1). An 80%/20% split was selected in order to hopefully allow the training sample to be sufficiently large (N=315) to form as good a model as possible, whilst having enough patients in the test sample (N=79), to be able to adequately assess the model fit. The final model coefficients are extracted from the training data model and used to predict the Larsen score for the test set. The correlation between the actual and predicted Larsen score is then plotted. The program used to produce this analysis is shown in Appendix G which invokes the same amended 'spls' and 'valid' functions as shown in Appendix D.

8.3.2. 'Average rank' method results using a separate training and test set

Using the 'average rank' method, a SPLS model containing the top ranked 100 variables was formed using the training set data. This model was used to predict the Larsen score for the independent test set. Unfortunately, the model revealed quite poor correlation, r=0.385 (Figure 8.6). No subject in this model was predicted with a Larsen score above 85 hence the full range of possible Larsen scores is not well represented by this model. However, generally this is not a problem with PLS methods as other models are able to predict the full range of values (Figure 6.11). In general, as all subjects are predicted a value between 10 and 85, subjects with a true value less than 10 are over-predicted and those greater than 85 are under predicted. Hence this model does not perform well at the extremes of the distribution.

In section 6.5.1 models containing from 5 to 505 variables were examined and a model with 100 variables was chosen because it revealed a good increase in correlation from 50 variables whilst hopefully not being over fitted (Figure 6.11). However, as using this model on an independent set appears to perform poorly, reassessment of whether the model is over fitted will be examined and how to select the optimum number of variables for the final model will be reinvestigated.

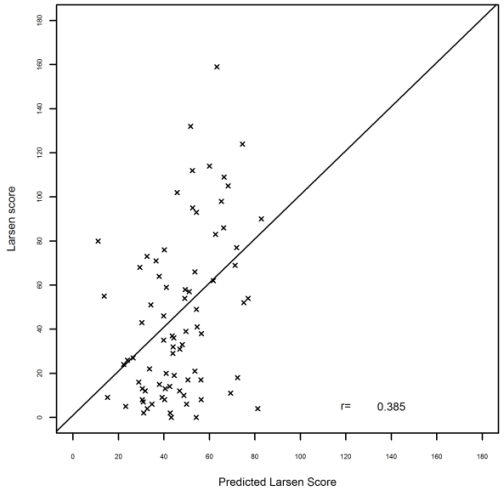


Figure 8.6 Independent prediction using the top 100 variables from the training model

8.3.3. 'Two stage average rank' method to select variables

The 'percentage fold' method uses internal CV to determine the optimal number of variables as recommended by González et al. (2011) and other authors as described in 4.2.1.4. Using the 'percentage fold' method on the 'all subjects' dataset (N=912, 368 SNPs) as shown in Figure 4.3 and Figure 4.6, the R^2-CV showed a peak and then a decrease in predictive ability when more variables were included in the model. It therefore appears to protect against over fitting, explaining why this method is often used in the literature. However, when using the 'percentage fold' method on the 'GWAS SNPs' dataset investigating 325482 SNPs (as shown in Figure 6.3), the model predicts increasingly better the more variables which are added. This is likely to be due to the high dimensionality of the data and has recently been reported by Le Floch et al. (2012) who suggest univariate filtering (or preferably multivariate filtering) as a 'mandatory step'. They warn that as SPLS (and PLS) attempt to explain all variation associated with the Y variable, the methods will not protect against the inclusion of multiple irrelevant predictors. It therefore appears that the success of the model to protect against over fitting decreases as the dimensionality of the data increases.

The 'average rank' method simply creates a list of variables in order of importance of their contribution to the model. Whilst the top 100 variables were used for the final model in section 8.3.2, this was only based on comparing the correlations from models with between 5 and 505 variables. A 100 variable model was selected as it 'appeared' to be a suitable increase in correlation for the number of variables added and hence may not be over fitting. Instead of investigating pre-filtering methods which are already being explored by authors (Le Cao et al., 2011, Le Floch et al., 2012), it was decided to investigate whether avoidance of over fitting could be obtained as part of the SPLS process.

To demonstrate the extent of the problem, the 'average rank' method was used on the 80% training set to create an order of variable importance. However, instead of using the top 100 variables (as used in section 8.3.2), the R^2-CV using training set only was calculated using models which contained from just two variables to 2000 variables. It was expected that as R^2 was calculated

under CV on the training set only, it would reveal a maximum number of variables after which adding further variables would only lead to over fitting of the model and hence lower R^2-CV. Figure 8.7 revealed a small increase from two variables (R^2-CV=0.56) to five variables (R^2-CV=0.59) after which it is relatively level until 100 variables (R^2-CV=0.59). However, there was an upward trend in R^2 from 100 to 500 variables (R^2-CV=0.616) and this trend continued until a R^2-CV=0.92 at 2000 variables. These results are in agreement with Le Floch et al,. (2012) who report in the high-dimensional setting, the more variables the model contains, the better the prediction. However, this is likely to be because the variables were chosen through an iterative process which selects variables based on their ability to predict well under CV and hence the model is still over fitted and would not perform well on an independent set.

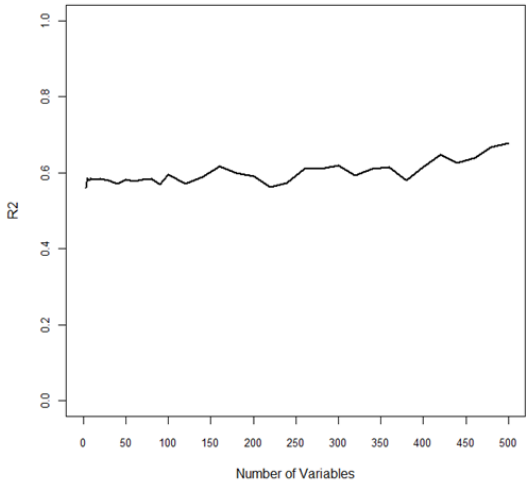


Figure 8.7 Using internal (80%) CV to determine optimum number of variables- 'GWAS SNPs' dataset

Whilst the 'average rank' method in section 6.5 removed the use of CV to select the optimum number of variables, the method resulted in a list of variables with no clear design of how to select the optimum for the final model. Internal CV (as shown in Figure 8.7) does not appear to protect against over fitting. It was therefore decided to investigate if using the 20% test set could determine the optimum number of variables to keep in the model instead. As the 20% test set is completely independent to the rest of the model fitting, it should still be able to control against over fitting. However, by using the test set to create the final model, it is no longer independent and as such cannot provide an unbiased estimate of model fit. For this reason the training set is referred to as the 'variable ordering training set' and the test set is referred to as the 'variable selection training set'.

A 'two stage average rank' method is defined as follows:

- Split the initial data into 80% of subjects for a 'variable ordering training set' and 20% of subjects for a 'variable selection training set'.
- Stage 1: The 'average rank' method is applied to the 80% (using 10 runs of 5-fold CV to average the ranked loadings for each variable) to create an ordered list of variables as described in section 6.4.1.
- Stage 2: From 2 to 100 variables are used from the ordered list and final models are created fixing the coefficients based on the 80% variable ordering training set. For example, the

first two variables from the ordered list are used to create a model estimating the coefficients for the two variables based on the 80% variable ordering training set only. This is the first model. Following on from this, the same process is used with the top three variables from the list and then four variables from the list continuing up to 100 variables. Each of these models are used to predict the 20% variable selection training set and the correlation between actual and predicted Larsen score is used to determine which model (with what number of variables) predicts the 20% of patients the best.

8.3.4. 'Two stage average rank' method results

Figure 8.8 reveals that the best prediction of the Larsen score for the 20% variable selection training set is when just three variables are used.

The model was: Larsen score = 43.15 + 0.745*DD + 0.664*SymDur - 0.501*agediag
Where DD= Disease duration (time since diagnosis in years), SymDur = Symptom duration (time since onset of symptoms in years) and agediag = age at time of diagnosis (years).

This resulted in a correlation between actual and predicted Larsen score of r=0.622 (Figure 8.9). Any additional variables included from the variable ordering training set model, lowered the correlation between the actual and predicted Larsen score on the variable selection training set. This suggested that the additional variables are over fitting the variable ordering training set model, explaining noise in the data and they are not predictive of the Larsen score when used on an independent set. It could also suggest that there is too much noise in the data to be able to determine the predictive signals, perhaps because the size of the signals are particularly small.

Figure 8.9 reveals that there is still substantial unexplained variability in the model. Despite this, the model containing just three variables performs substantially better at the upper extremes of the Larsen score than the single stage 'Average rank' method (Figure 8.6) which contained 100 variables. For example, the maximum value predicted is almost 120 and three subjects between 100 and 120 are predicted almost exactly. However, subjects with lower Larsen scores (particularly less than 20) are consistently over-predicted. When the Larsen score is plotted against disease duration (Figure 2.2), it was clear numerous subjects did not develop the severity of erosions expected based on their disease duration. Therefore, as this model does not contain any genetics or environmental variables, it is likely some key variables which explain why some subjects develop more severe disease than others are missing from the analysis. Unfortunately, with such a small sample of patients (training set=315, test set=79), the model is not able to reliably identify these variables.

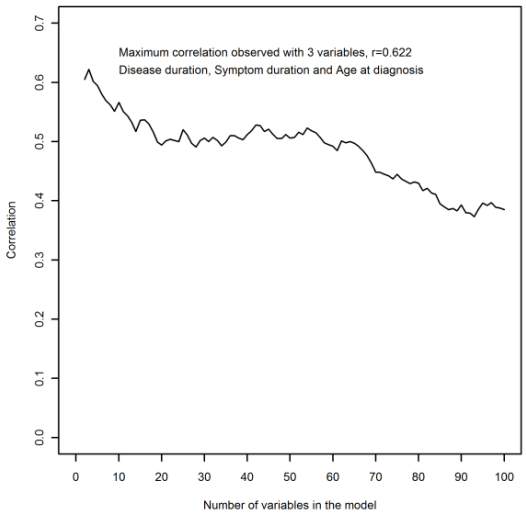


Figure 8.8 Correlation between actual and predicted Larsen score for 2 to 100 variables

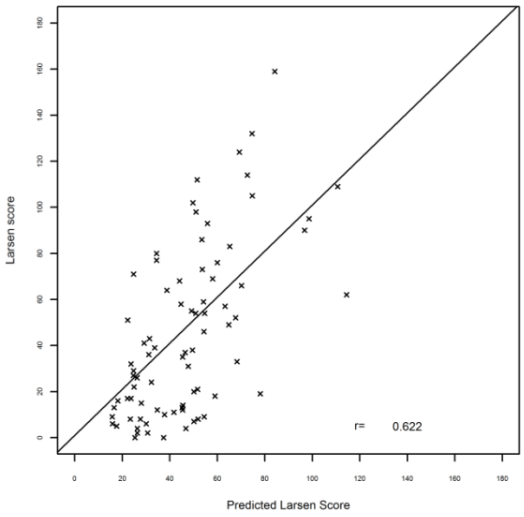


Figure 8.9 Independent prediction using the top three variables from the training model

8.3.5. Summary of results using 'average rank' methods on 'GWAS SNPs' dataset

In conclusion, if the variable ordering has been performed using internal CV of a training set, CV cannot be then used again with the same subjects to determine the optimum number of variables for the model. Although pre-filtering could be used as suggested by Le Floch et al. (2012), Abraham et al. (2013) suggest in other penalised regression methods that this may lead to a reduced ability to detect causal SNPs.

Therefore, a solution was suggested entitled the 'two stage average rank' method, which used 80% of the data to create the ordered list of variables and 20% to select what number of variables to include in order to avoid over fitting. However, by using the independent sample in this way to estimate the number of variables required, this leads to overestimation of how well the model would perform on an independent set.

Using the 'two stage average rank' method, the best estimate of the Larsen score is calculated simply by: Larsen score = 43.15 + 0.745*DD + 0.664*SymDur - 0.501*agediag
Where DD= Disease duration (time since diagnosis in years), SymDur = Symptom duration (time since onset of symptoms in years) and agediag = age at time of diagnosis (years). Hence with such a small sample of patients, SPLS modelling was not able to identify any SNPs or environmental factors contributing to RA severity.

8.4. Independent training and test sets – 'All subjects' dataset

8.4.1. 'Average rank' method using a separate test and training set

In order to have more subjects available for use in the modelling, the 'average rank' method (on separate 80% training and 20% test sets) is applied to the 'all subjects' dataset. Although this reduces the number of SNPs available to 368, it may give insight as to whether a better model could be determined if there were more subjects.

Sorting the dataset by the Larsen score, every 5th subject is reserved for the test set (which is used for training variable selection in the 'two stage average rank' method). This results in 730 subjects in the training set and 182 in the test set (80% training, 20% test). Ten runs of 7-fold CV are used so that approximately 104 subjects are left out of the training set for each of the seven folds. 7-fold CV is used instead of 10-fold CV (section 4.4) because it was the SIMCA default (observed in chapter 7) and there are 182 fewer subjects in the model creation (training) dataset. As there are fewer SNPs being investigated, the 'all subjects' dataset does not require PLS models to be fitted hierarchically (extracting the top 200 variables for each chromosome and then fitting a higher level model).

The 'average rank' method is used to produce an order of predictive importance for each of the variables based on the 80% training dataset. Instead of using the top 100 variables for the final model which was used in section 8.3.2 on the 'GWAS SNPs' dataset, it was decided to reapply a technique based on González et al. (2011) detailed in section 4.2.1.4, in an attempt to avoid over fitting of the model. This may be a more applicable approach on the 'all subjects' dataset since the hierarchical approach (multiple CV models) to select variables is not applied to this smaller set of data. After estimating the optimum number of variables for the final model based on the 80% training dataset, the model with fixed coefficients is then used to predict the 20% test set.

8.4.2. 'Average rank' method results using a separate test and training set

The 'average rank' method was applied to the 'all subjects' dataset and an order of predictive importance created. To investigate the optimum number of variables to include in the final model, one run of 7-fold CV was fitted to models containing from two to 250 variables. Each models R^2-CV was plotted against the number of variables that model contained. The maximum R^2-CV was observed when the model contained 95 variables (R^2-CV=0.5811) however the maximum was almost reached at 70 variables (R^2-CV= 0.5794) (Figure 8.10). It was therefore decided to base the final model for prediction of the test set on the top 70 variables from the training set model.

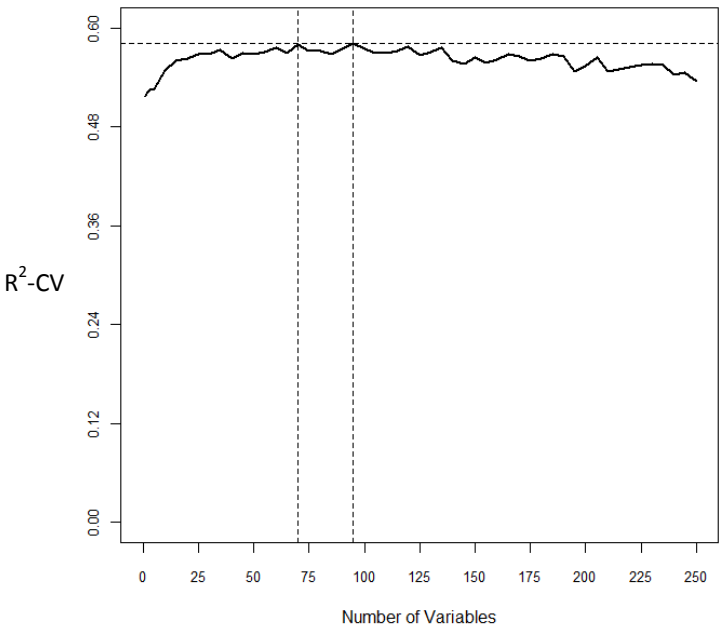


Figure 8.10 Determination of the optimum number of variables to retain for the final model – 'All subjects' dataset

The order of importance of the variables was similar to section 4.4, however, as ACPA and BMI were now included as environmental variables (included as described in section 6.2.2), they featured high in the order of importance. Using the top 70 variables to predict the Larsen score on the independent test set resulted in a correlation between actual and predicted Larsen score of r=0.456.

8.4.3. 'Two stage average rank' method

In order to investigate whether the correlation of r=0.456 could be improved by using a different number of variables to the 70 selected in 8.4.2, it was decided to apply the 'two stage average rank' method (as described section 8.3.3). The same 80%/20% splits are used as described in 8.4.1. If improved model prediction can be obtained with fewer variables then it suggests the model is over fitted. This would indicate that CV is not appropriate to determine the number of variables even in the case of the smaller 'all subjects' dataset.

8.4.4. 'Two stage average rank' method results

When the model coefficients are formed on the 80% variable ordering training set and used to predict the 20% variable selection training set, modelling revealed that the number of variables required to have in the model, which achieved the best prediction, was seven (Figure 8.11). Just these seven variables resulted in a higher correlation (r=0.576) than using 70 variable in section 8.4.2.

These findings are supportive of section 8.3.4 concluding that determining the number of variables to retain in the final model based on internal CV, does not protect the model against over fitting. Even on the smaller 'all subjects' dataset, the model is still considerably over fitted when the number of variables to include is estimated using CV as recommended by (González et al., 2011).

The top seven variables were disease duration, symptom duration, age at onset of symptoms, age at time of diagnosis, ACPA category, BMI and ACPA value (Figure 8.11). There was wide prediction error particularly for subjects with Larsen score equal to zero or very high Larsen score (Figure 8.12). Similar to 8.3.2, this model had poor prediction at the extremes of the Larsen score distribution, with no predicted Larsen score result over 85. This is probably due to having very few variables included in the model, resulting in the wide variation in the Larsen score not being able to be predicted.

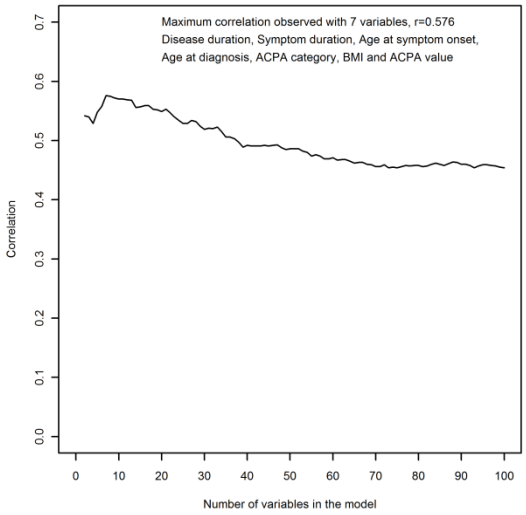


Figure 8.11 Correlation between actual and predicted Larsen score retaining two to 100 variables – 'All subjects' dataset.

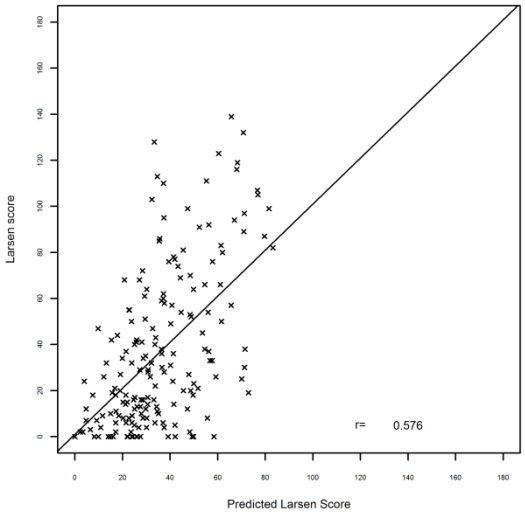


Figure 8.12 Independent prediction using the top seven variables from the training model – 'All subjects' dataset.

8.4.5. 'Two stage average rank' method for a two component model

Le Floch et al. (2012) identified the SPLS tendency to retain too many variables contributing to the components resulting in an over fitted model. Based on this, it was considered possible that the dataset being explored is so large that truly important predictors which would normally be found contributing to a second component, are being overshadowed by multiple predictors (not truly predictive) but successfully contributing to the R^2 on the first component.

As described in 3.7.1, the residual error which represents the amount of variation left to explain in the X and Y data is calculated after each component's scores (t_h) and loadings (p_h) have been calculated. This can be written for the X variables as $E_h = E_{h-1} - t_h p'_h$; $X=E_0$ and for the Y variable(s) as $F_h = F_{h-1} - b_h t_h q'_h$; $Y=F_0$ where h is the number of components. Therefore, if too many variables are included in the first component such that it creates an over fitted model explaining much of the variation in the data, there will be little residual error left to be predicted by the second component (and subsequent components). However, if a method can restrict the first component to only variables truly predictive, then a better model may be achieved using two or more components.

This may explain why no model has required two components in any of the work to date. A review of the literature revealed that most authors appear to apply pre-filtering of SNPs and hence the problem does not appear to have been explored.

It was decided to use the 'two stage average rank' method on the 'all subjects' dataset to see if a more successful model can be created with more than one component. The number of variables for the first component will be restricted, based on the 'two stage average rank' method results from section 8.4.3 (seven variable model protected against over fitting) and then the same 'two stage average rank' method will be refitted to derive a second component. By using the independent validation to prevent over-fitting in the 1st component, there may be sufficient variation unexplained in the Larsen score for a 2nd component to be beneficial to the model prediction. See section 9.2 for further exploration into the use of two components in the context of multiple Y severity variable modelling.

8.4.6. 'Two stage average rank' method results for a two component model

In accordance with section 8.4.3, a model containing seven variables for the first component (disease duration, symptom duration, age at onset of symptoms, age at time of diagnosis, ACPA category, BMI and ACPA value) was fitted using the training dataset. The residual variation which could not be explained by the first component was then used to fit a second component and the variables were ordered according to the average absolute size of the loading vector corresponding to the 2nd component across 10 runs of 7-fold CV.

The seven variables for the first component and from two to 100 of the top ranked variables for the second component were fitted and the coefficients for the model obtained based on the training data. These models were then used to predict the Larsen score for the test set. The highest correlation (r=0.579) between the actual and predicted Larsen score was obtained when the top ten variables were retained for the 2nd component. This equated to 14 variables in the model in total, as there were seven in 1st component and 10 in 2nd component with three overlapping both components (Figure 8.13). The 14 variables were: disease duration, symptom duration, age at onset of symptoms, age at time of diagnosis, ACPA category, BMI, ACPA value, rs9366826, age, alcohol use, rs805292, rs26232, rs394581 and rs2075800. Figure 8.14 shows the correlation (r=0.579) for the two component model between the actual and predicted Larsen score.

A correlation of r=0.576 was obtained using one component and seven variables. Therefore, it appears the second component adds very little extra predictive ability (increase in correlation of 0.003) even when the 1st component is restricted to very few variables. In this example, it appears that the initial tests commonly used with PLS for inclusion of additional components (described in 4.2.1.3) are adequate. This will be explored again in chapter 9, when multiple Y variables may make additional components contribute to improved prediction in the model.

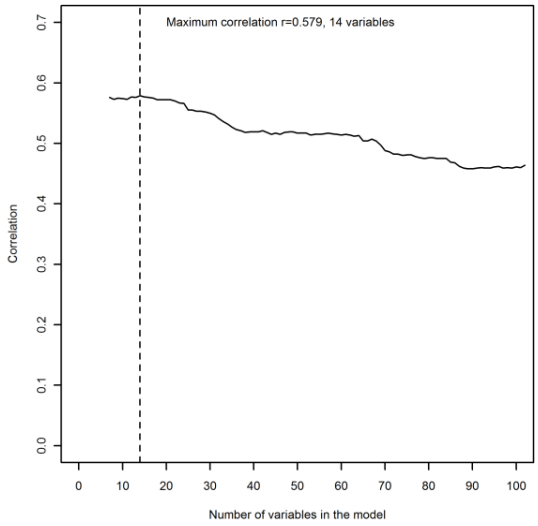


Figure 8.13 Correlation between actual and predicted Larsen score retaining two to 100 variables for the 2nd component– 'All subjects' dataset

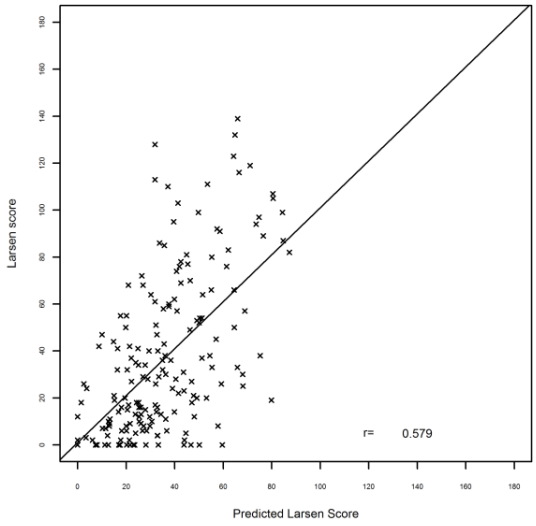


Figure 8.14 Independent prediction using the top 14 variables and two components from the training model – 'All subjects' dataset

8.4.7. 'Three stage average rank' method

A fundamental problem with the 'two stage average rank' method is that as the 20% independent test sample is used to determine the number of variables in the second stage, it is no longer independent from the model creation and cannot provide an independent estimate of model performance.

Therefore a 'three stage average rank' method was developed and is described below:

- Split the initial data into 40% of subjects for 'variable ordering training set' and 40% of subjects for 'variable selection training set' and 20% of subjects for independent test set.
- Stage 1: The 'average rank' method is applied to the 40% 'variable ordering training set' (using 10 runs of 7-fold CV to average the ranked loadings for each variable) to create an ordered list of variables.
- Stage 2: From 2 to 100 variables are used from the ordered list and final models are created fixing the coefficients based on the 40% 'variable ordering training set' (as described in section 8.3.3). The Larsen score of the subjects in the 40% 'variable selection training set' are predicted using all of the models and the model with the highest correlation between actual and predicted Larsen score selected as the final model.
- Stage 3: Using the final model chosen in stage 2 (with a fixed number of variables and fixed coefficients) the 20% independent test set subjects are predicted. As these subjects were not used at all in creation of the model it should better represent the model performance on an independent sample.

Data was split into three sets containing 40%, 40% and 20% using the ordered Larsen score. After ordering the patients by the Larsen score, patient 1 was assigned to group 1, 2 to group 2 and so on

until patient 10 was assigned to group 10. Patient 11 was then assigned to group 1, 12 to group 2 and so on until all patients were assigned to one of the 10 groups. Groups 1, 5, 6 and 10 were assigned to the 40% 'variable ordering training set' (N=365), groups 2, 4, 7 and 9 were assigned to the 40% 'variable selection training set' (N=365) and groups 3 and 8 were assigned to the independent 20% test set (N=182). This ensures a similar spread of Larsen score results in each of the sets of data.

8.4.8. 'Three stage average rank' method results

Stage 1 of the 'three stage average rank' method was executed using the 40% 'variable ordering training set'. A list of variables according to importance was created. Creating models on the variable ordering training set containing from 2 to 100 variables and testing it on the 40% 'variable selection training set' resulted in the highest correlation between actual and predicted Larsen score being achieved with 10 variables (r=0.604, Figure 8.15). The 10 variables were: Disease duration, symptom duration, age at symptom onset, Age at diagnosis, ACPA category, ACPA value, rs26510, BMI, DRB1 S2 and rs26232.

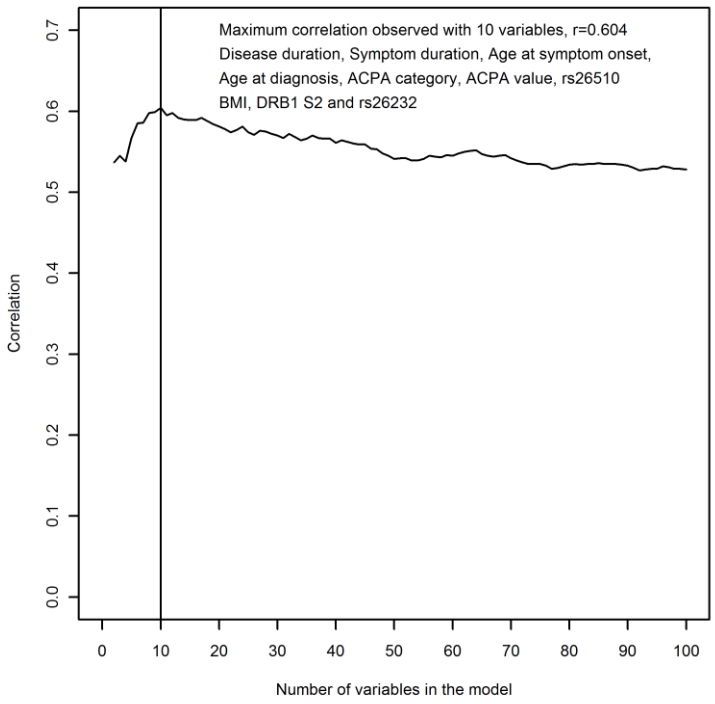


Figure 8.15 Stage 2 of the 'three stage average rank' method applied to the 'all subjects' dataset

The final model was:
Larsen score= 50.33 + 0.543*DD + 0.515*SD - 0.290*Ageonset - 0.283*Agediag + 6.445*ACPA Category + 0.051*ACPA value - 3.161*rs26510 - 0.43*BMI + 3.131* DRB1 S2 - 2.964*rs26232

Where DD= Disease duration (time since diagnosis in years), SD = Symptom duration (time since onset of symptoms in years), Ageonset = Age at symptom onset (years), agediag = age at time of diagnosis (years), ACPA category (coded as a 1=positive, 0=negative), rs26510 and rs26232 coded according to the frequency of the minor allele (0,1,2), BMI (in kg/m^2) and DRB1 S2 (according to the

number of alleles (0, 1 or 2) with the amino acid sequence of K-R-A-A motif at positions 71-74 of the HLA-DRB1 region in the third hypervariable region of the DR molecule).

For the first time in this research, genetic variants have been identified as predictive on an independent test set. DRB1 S2 and rs26232 (C5orf30) were previously identified in the literature review in section 2.3.2. rs26510 is a C/T polymorphism located on chromosome 5 at position 96125910. It is located in the intron region of the endoplasmic reticulum aminopeptidase 1 (ERAP-1) gene. No previous links to RA disease were found in the literature.

Stage 3 was then performed using this model to predict the 20% of subjects in the independent test set. A correlation of 0.559 was achieved (Figure 8.16) which is a substantial improvement compared to using the simpler one stage 'average rank' method model (r=0.456) when it was tested on an independent set using 70 variables in section 8.4.2. However, the model still suffers from an inability to predict very high and very low Larsen score values with the maximum predicted value being approximately 90. There is wide variation in the prediction ability across all Larsen score values. Therefore, although the use of 10 variables is protecting against over fitting, there is substantial unexplainable variation and potentially key variables missing from the model.

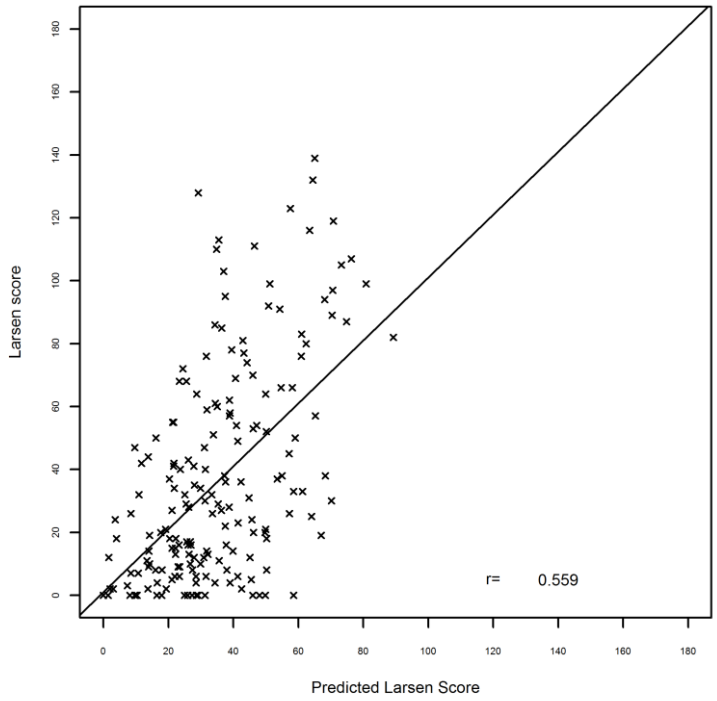


Figure 8.16 Stage 3 (independent prediction) using the 'three stage average rank' method applied to 'all subjects' dataset

8.4.9. Summary of results using 'average rank' methods on 'All subjects' dataset

This section found evidence that over fitting of SPLS models is still a problem even when CV is used on a relatively small dataset (912 subjects, 368 SNPs and 19 environmental variables). Whilst in section 8.3.5, a two stage procedure appeared to reduce the risk of over fitting, it resulted in no

independent dataset available to test the final model on (one which is not involved at all in the model creation). One solution to this problem is to split the data into three different sets. The first set is used in stage 1 to create a list of ordered variables, the second set is used in stage 2 using independent testing to select the number of variables to keep for the final model and the third set is used in stage 3 to provide an estimate of the prediction ability of the model under independent testing.

The resulting model consisting of Disease duration, symptom duration, age at symptom onset, Age at diagnosis, ACPA category, ACPA value, rs26510, BMI, DRB1 S2 and rs26232 was able to predict 182 independent patients (not used in model creation) with a correlation of r=0.559. This was the first model developed in this research to include genetic components when tested independently of a separate set of patients not used in the model fitting.

All models explored in section 8.3 and 8.4 tended to over-estimate the predicted Larsen score for subjects with low Larsen score and under-estimate it for subjects with high Larsen score. As this was not a problem when the model was over-fitted (section 6, Figure 6.9), it is believed to be due to the model not including variables which can explain variation in the extremes of the Larsen score distribution. For example, if two patients have identical disease duration, environment and genetics, however experience different treatment regimens, the treatment regimen could be responsible for observing different severities. Unfortunately, details of treatments administered to subjects over previous years are not available for the GoRA subjects.

8.5. 'Three stage average rank' method - Subset analysis

The 'three stage average rank' method developed in section 8.4.7 was applied to three subsets of disease to investigate whether improved model performance could be achieved by reducing the heterogeneity of the cohort.

8.5.1. Disease duration subsets

8.5.1.1. Methods

It is hypothesised that the wide prediction error may be caused by patients having varied disease duration. Hence, factors influencing disease severity over many years, which were not measured in the cross-sectional study, cannot be accounted for. In order to investigate further, subsets of the 'all subjects' dataset were created for subjects with disease duration less than 10 years and less than 15 years. Ideally, an even more homogenous sample would be created using patients with disease duration less than three or maybe five years disease duration, however, the sample size is insufficient to use a cut off any less than 10 years. Ten and 15 years is therefore chosen to provide enough patients for the analysis whilst still limiting subjects to those who have received relatively modern and hopefully more similar treatment regimens.

350 subjects had disease duration less than 10 years. This data is split into 40% for a 'variable ordering training set' (N=140), 40% for a 'variable selection training set' (N=140) and 20% for an independent test set (N=70), in order to apply the 'three stage average rank' method described in 8.4.7.

535 subjects had disease duration less than 15 years which are split into 40% for a 'variable ordering training set' (N=214), 40% for a 'variable selection training set' (N=214) and 20% for an independent test set (N=107), in order to apply the 'three stage average rank' method described in 8.4.7.

8.5.1.2. 'Three stage average rank' method results: disease duration < 10 years subset

Figure 8.17 revealed that the maximum correlation which could be achieved creating a model using the 40% variable ordering training set and testing it on the 40% variable selection training set was observed using the top 6 variables (r=0.364). The final model consisted of Larsen score = 22.7 + 2.193*DD - 0.757*BMI + 5.348*rs443198 + 7.218*rs2568127 + 9.022*rs4133002 - 5.79 *rs4535211.

Where DD= Disease duration (time since diagnosis in years) and BMI (in kg/m^2).

This model was then used to predict the independent test set. The correlation between actual and predicted Larsen score for the 20% independent test was r=0.284 (Figure 8.18). Therefore, reducing the sample to just those subjects <10 years disease duration does not increase model prediction ability.

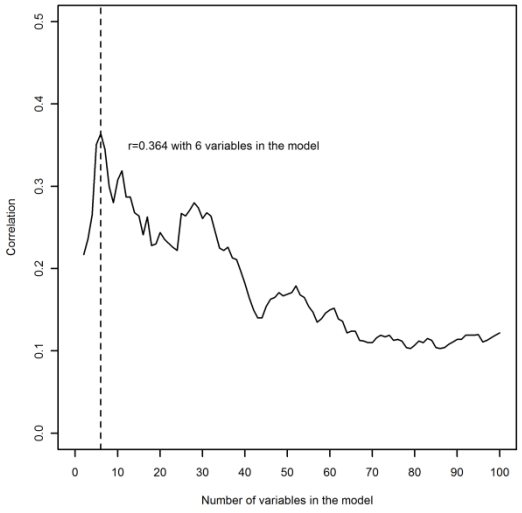


Figure 8.17 Stage 2: Correlation between actual and predicted Larsen score retaining two to 100 variables – Disease duration <10 years

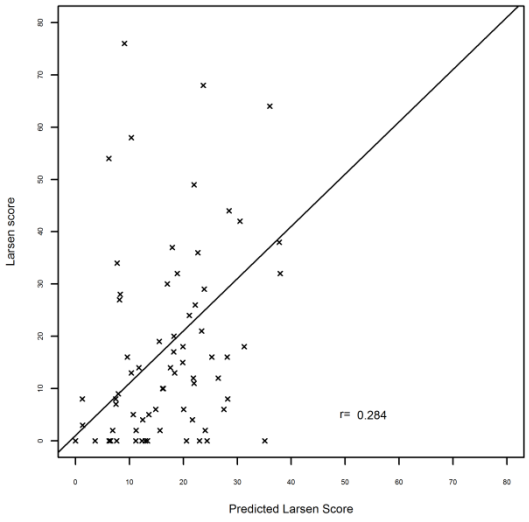


Figure 8.18 Stage 3: Independent prediction using the top six variables from the training model – Disease duration <10 years

8.5.1.3. 'Three stage average rank' method results: disease duration <15 years subset

Based on the subset of patients with disease duration less than 15 years, the maximum correlation was obtained with just two variables (r=0.373, Figure 8.19). The model consisted of Larsen score = 3.16709 + 2.264* DD + 11.357*ACPA category. Where DD= Disease duration (time since diagnosis in years) and ACPA category = 1: positive, 0=negative.

Using the above model to predict the 20% independent test set, achieves a correlation of r=0.424. However, as can be seen in Figure 8.20 the maximum predicted score is less than 50 which is very

low compared to the maximum actual Larsen score which is 112. Therefore the model performs poorly, over predicting the lower values and under predicting the higher values.

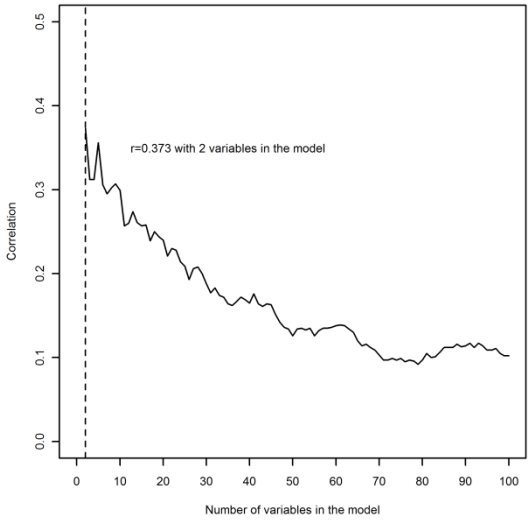


Figure 8.19 Stage 2: Correlation between actual and predicted Larsen score retaining two to 100 variables – Disease duration <15 years

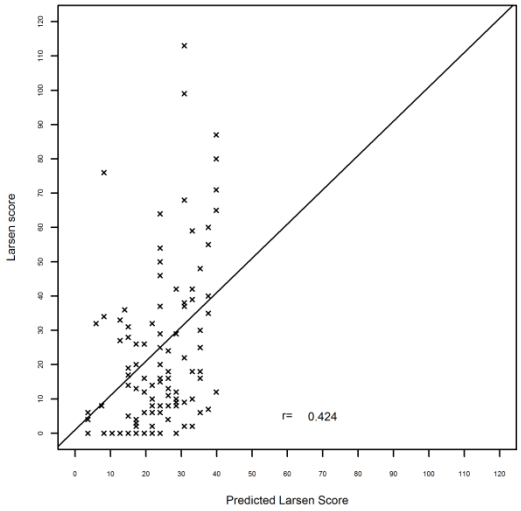


Figure 8.20 Stage 3: Independent prediction using the top two variables from the training model – Disease duration <15 years

8.5.2. ACPA positive subset

8.5.2.1. Methods

ACPA is thought to categorise patients into two distinct subsets of disease as described in section 6.2.2. Genetic variants have previously been observed to have a different association with severity depending on whether the subject is ACPA positive or ACPA negative (sections 2.3.2.5 and 2.3.2.6). Although ACPA was included as a variable in the model, very little research has been performed investigating the ability of PLS to model interactions as noted in section 3.7.3. Therefore in this section, ACPA positive patients are analysed on their own to determine whether there is any increase in predictive ability of the model using a more homogenous dataset. ACPA negative will not be investigated as the sample size (N=223) is too small to split into a training and test sample.

689 subjects in the GoRA cohort are ACPA positive which were split into 40% of subjects for a 'variable ordering training set' (N=275), 40% of subjects for a 'variable selection training set' (N=276) and 20% subjects for an independent test set (N=138). Models were fitted per the 'three stage average rank' method described in section 8.4.7.

The variables ACPA category and ACPA value were excluded from this modelling.

8.5.2.2. 'Three stage average rank' method results: APCA positive subset

Stage 2 of the 'average rank' method suggested that a model containing the top six variables could predict the 'variable selection training set' with the highest correlation (r=0.629, Figure 8.21). The model consisted of: Larsen score =62.58 + 0.574* DD + 0.508 * SD – 0.295 *ageatdiag – 0.288*ageatonset – 0.457*BMI – 4.937*rs2073839. rs2073839 is a intron variant in the solute carrier family 22 (SLC22) A4 gene on chromosome 5.

Using this model to predict the 20% independent test set resulted in a correlation of r=0.611 (Figure 8.22). This was a slight improvement to modelling the 'all subjects' dataset using the 'three stage average rank' method which resulted in a correlation of r=0.559. It is possible that this slight increase in correlation could be due to random variation.

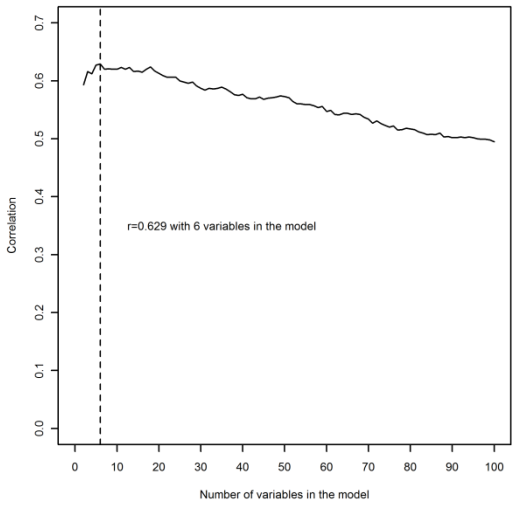


Figure 8.21 Stage 2: Correlation between actual and predicted Larsen score retaining two to 100 variables – ACPA positive subjects

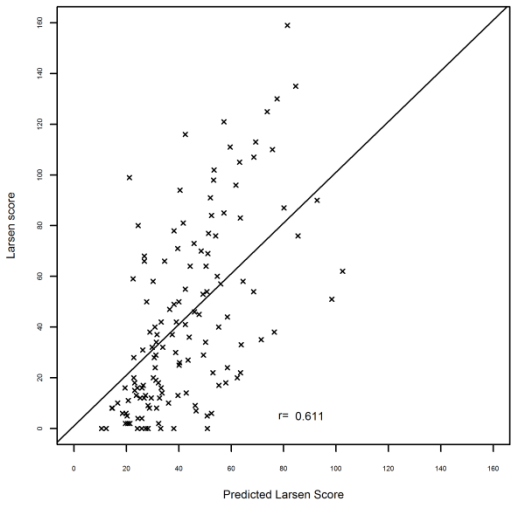


Figure 8.22 Stage 3: Independent prediction using the top six variables from the training model – ACPA positive subjects

8.6. Summary

This chapter focused on validation methods to attempt to quantify whether the models created are over fitted and how models may perform on an independent cohort. A new technique was developed to prevent against over fitting and provide an estimate of independent prediction entitled the 'three stage average rank' method.

Initially, a permutation approach was adopted to investigate the ability of the 325,482 SNPs plus environmental variables to predict the Larsen score, under the null hypothesis that there is no relationship. The Larsen score was randomly permuted 100 times and the full modelling process (developed in chapter 6) repeated. The proportion of total Larsen score variation explained by the model fitted, was calculated under CV (R^2-CV) for the 100 permuted datasets and compared to the R^2-CV for the true Larsen score model. The true Larsen score model had a higher R^2-CV (0.753) than could be achieved by any of the randomly permuted Larsen score datasets (section 8.2). This suggested that it is highly unlikely to achieve this result by chance ($p<0.01$), if no relationship

existed between the X and Y variables. Following this analysis, a further permutation analysis was performed retaining the link between the Larsen score and environmental variables whilst removing the link between the Larsen score and the SNPs. This investigated whether any additional predictive ability after removing that associated with the environmental variables is due to the SNPs. Two of the 100 permutations were found to achieve an R^2 greater than the true Larsen score model. This indicates that it is unlikely (p=0.02) to have a model which can predict the true Larsen score with an R^2 of 0.866, if the SNPs contribution is due to chance alone.

The 'GWAS SNPs' dataset was split into two datasets (80% for training and 20% for test) based on the distribution of the Larsen score to ensure fair representation of all possible values in the test set. The SPLS 'average rank' method was produced on the 80% training set, with variables forming part of the higher level model if the median (across the folds and runs) of their ranked loadings were less than 200 after 10 runs of 5-fold CV. 10 runs of 5-fold CV were re-run on the higher level model to produce a final ranked order of the variables. Investigation at this stage under CV revealed that it was unable to estimate the minimum number of variables required in order to prevent over fitting of the model. Instead, it suggested that the more variables which were included (right up to 2000 variables) improved the model fit. To reflect the model created in section 6.5.2, the top 100 variables were chosen and the model coefficients calculated using the training set. This model was then used to predict the test set, however resulted in very poor correlation (r=0.385). The modelling process appeared unable to protect against over fitting.

Therefore, a 'two stage average rank' method was developed which attempted to select the optimum number of variables to retain in order to avoid over fitting. The order of variables determined by the 'average rank' method was used to investigate models containing from two to 100 variables. For each of these models separately, the coefficients were fixed using only training data before predicting the test set. This revealed that using disease duration, symptom duration and age at time of diagnosis outperformed any model using more variables (r=0.622). This suggests that the SPLS modelling approach with this cohort could not detect any genetic variants which improved the model fit on an independent dataset. Unfortunately, because the test set had to be used to determine the best number of variables for the final model, it is no longer independent and could be over exaggerating performance on a truly independent set.

The SPLS approach, using the 'GWAS SNPs' dataset with a relatively small sample size (compared to the number of X variables) and widely varied disease duration, was unable to find any genetic variables which could add any benefit to the prediction of the Larsen score on an independent dataset. Therefore, it was decided to increase the sample size (and reduce the number of X variables) by investigating the 'all subjects' dataset with only 368 SNPs.

Using the 'average rank' method (but without the hierarchical modelling for the 40 blocks of chromosomes), the 'all subjects' dataset was split into a training set (N=730) and a test set (N=182). It was decided to use 7-fold CV as this was the default in SIMCA and there were only 730 subjects in the training set instead of the full 912 previously modelled when 10-fold CV was used.

Initially, to decide how many variables to extract from the 'average rank' method models, 7-fold CV was run once, extracting a different number of variables each time and estimating the R^2-CV. Seventy variables were found to provide approximately the optimum R^2-CV. Unlike the GWAS modelling, the graph produced a clear peak before descending, suggesting that the reason this

process doesn't work in the GWAS modelling is because the variables have already been chosen, using the same subjects in the lower level models under CV.

A model containing 70 variables (with fixed coefficients based on the training set) was used to predict the Larsen scores for patients in the test set. It resulted in a correlation between the actual and predicted Larsen score of r=0.456. To determine whether the model was still over fitted or whether CV to determine the number of variables was protecting against this, a 'two stage average rank' method was developed.

After using the 'average rank' method to provide an order of importance for the variables, separate models containing from two to 100 variables were created fixing the estimates of the coefficients using only the training data. These models were then used to predict the Larsen scores for patients in the test set. The maximum correlation (r=0.576) was achieved with just seven variables (disease duration, symptom duration, age at onset of symptoms, age at time of diagnosis, ACPA category, BMI and ACPA value). Hence, determining the number of variables to retain in the final model based on internal CV does not protect against over fitting. Authors have recently identified this issue using other penalised regression methods, particularly when the sample size is not large enough for the number of variables being modelled (Li and Sillanpaa, 2012, Ayers and Cordell, 2010, Abraham et al., 2013). In addition, the issue has also been identified as a problem in PLS with the only solution recommended being the pre-filtering of SNPs (Le Floch et al., 2012, Le Cao et al., 2011). Reassuringly, the model using the two stage process selected variables previously identified in the literature review (section 2.4) as key environmental predictors of RA severity. However, the model was not able to identify any genetic variants predictive of the Larsen score.

The addition of a second component was explored to investigate whether over fitting in the first component is the reason a second component is not required. However the second component was found to be unnecessary even when the first component was restricted to just the top seven variables, providing evidence that the method of choosing the number of components is adequate.

To enable independent testing of the 'two stage average rank' method, a three stage method was developed. Data was split into a 40% sample to produce an ordered list of importance of the variables using the 'average rank' method. Models were then fitted including various numbers of the ordered predictive variables, fixing the coefficients using the initial 40% of data and predicting the independent 40% of data. Once the optimum number of variables was selected, the model was used to predict the remaining independent 20% of subjects.

This new method, was able to predict the independent set of 182 subjects with a correlation between the actual and predicted Larsen score of r=0.559, using a model formed on training data containing 10 variables. The 10 variables were: Disease duration, symptom duration, age at symptom onset, Age at diagnosis, ACPA category, ACPA value, rs26510, BMI, DRB1 S2 and rs26232. All of which have been found in the literature review to be predictive of RA severity with the exception of rs26510, which is a C/T polymorphism located on chromosome 5 in the intron region of ERAP-1.

In conclusion, the 'three stage average rank' method allows variable ordering, variable selection and independent testing to assess model performance. However, the initial sample has to be quite large to have sufficient patients to split the data into 40% variable ordering, 40% variable selection

and 20% model testing datasets. For this reason and due to time constraints, the method was not applied to the 'GWAS SNPs' dataset.

Patients were grouped into <10 years disease duration, <15 years disease duration and ACPA positive subjects, to attempt to reduce heterogeneity (perhaps due to unmeasured treatment regimen effects). Using the 'three stage average rank' method, limiting the sample to a duration of disease of <10 years (N=350) or <15 years (N=535) did not improve the model fit compared to using the 'all subjects' dataset (r=0.284 and r=0.424 respectively). One reason for this, particularly in the <10 years set could be due to insufficient sample size in the training data to form a good model. The evidence for this was that only disease duration and BMI were selected from the environmental variables to be included in the model, whereas most other models included variables such as symptom duration, age at symptom onset, age at disease diagnosis or the ACPA variables.

The ACPA positive patients subset (N=689), using the 'three stage average rank' method, achieved a correlation between the predicted and actual Larsen score of r=0.611. This was achieved using a model containing the following six variables (disease duration, symptom duration, age at onset of symptoms, age at time of diagnosis, BMI and rs2073839). rs2073839 is a previously investigated SNP for RA susceptibility which is found in the intron region of the solute carrier family 22 (SLC22) A4 gene in chromosome 5. This model achieved a slightly better prediction than using the 'all subjects' dataset (r=0.611 versus r=0.559) although variability estimates were not explored.

In summary, the 'GWAS SNPs' dataset using the 'two stage average rank' method could not produce a Larsen score predictive model which performs any better than using disease duration, symptom duration and age at time of diagnosis (r=0.622). The 'all subjects' dataset using the 'three stage average rank' method was able to create a model capable of predicting an independent subset of the GoRA cohort with a correlation of r=0.559 using the following ten variables; Disease duration, symptom duration, age at symptom onset, Age at diagnosis, ACPA category, ACPA value, rs26510, BMI, DRB1 S2 and rs26232. The predictive correlation was improved by modelling the ACPA positive patients on their own using a model with the following six variables (disease duration, symptom duration, age at onset of symptoms, age at time of diagnosis, BMI and rs2073839), r=0.611.

Variability estimates for the correlations were not calculated, which has recently been recommended by Daetwyler et al. (2013). This would form the basis for further research as described in section 10.4.

9. SPLS regression of multiple RA severity measures

9.1. Aims

Using lessons learned from modelling a single Larsen score variable, the aim of this chapter is to investigate whether modelling multiple RA severity measures together can lead to any benefit in predicting the severity of RA.

9.2. Methods

Although to date, only the single Larsen score Y variable has been predicted by the modelling, the GoRA cohort has substantial other information about the severity of the patient's RA. Eriksson et al. (2006a, p. 23) argue that "All data points are needed". They advise against selecting just the best variable or analysing one variable at a time. Instead they promote the analysis of multiple collinear variables together which optimises the use of the information. For this reason, this chapter models multiple correlated RA severity measures. Three groups of severity variables are investigated as defined in section 2.2.1.2.

1) 8 domain SF-36 variables
2) 4 SJC/TJC variables & DAS28
3) PVAS, RASEV, MHAQ, ESR, CRP, Any Erosions, the Larsen score (including the separate hand and foot counts) & DAS28

Section 8.4 ('all subjects' dataset) was able to achieve a better correlation between the predicted and actual Larsen score than section 8.3 ('GWAS SNPs' dataset). Therefore, it was decided to use the 'all subjects' dataset to model the multiple RA severity measures. Instead of 912 subjects modelled in chapter 4 and section 8.4, 914 subjects can be included. This is because, although two subjects had missing Larsen score data, they have other severity measures present. The dataset consists of 398 SNPs, five DRB1 variables coded according to Tezenas du Montcel et al. (2005), 19 environmental variables and 22 severity variables.

The mixOmics functions in R required further updates (as shown in Appendix H) to enable suitable output data when modelling multiple Y variables and multiple components instead of using SIMCA. Although SIMCA is substantially quicker when analysing very large datasets, the variable selection aspects of model fitting and splitting the sample into a training and test set outside of the CV cannot be automated. Therefore, it can be slower when analysing a relatively small dataset and subjects are required to be selected for a separate training and independent test samples (section 7.4). The programming code that invokes the functions shown in Appendix H, which were used to produce these analyses, is shown in Appendix I.

A SPLS analyses using the 'two stage average rank' method, is performed for each of the three multiple Y variable severity groups, using 80% for the variable ordering training set (N=731) and 20% as the variable selection training set (N=183) (independent test set). In order to get a fair representation of the distribution of the severity variables in two sets, subjects are sorted by a severity variable representative of the group and every 5th subject starting with the third subject is selected to be in the test set. As the groups are defined by variables which are correlated together, the mean of all of the 8 domains of the SF-36 is used to sort the subjects for analysis one and the DAS28 variable is used to sort the subjects for analyses two and three. As DAS28 was a composite

measure it was anticipated this would represent the other variables due to the correlation observed in Figure 2.3.

NIPALS imputation is performed in R prior to model fitting. Variables are checked to ensure they had sufficient variation in the training data during the CV. Any variables with >92% all 0's were excluded as there was insufficient variation for the model to fit.

As the 'average rank' method was new to this research, no prior strategy to enable selection of variables with more than one component is available. Le Cao et al. (2008) use the loadings from each component to determine which variables contribute to that component, based on the number of times the variable is selected in each of the folds and runs of the model ('percentage fold' method). Hence different variables can contribute to different components.

Whilst the ranking of the variables could be selected based on the average or the sum of the loadings across the components, this would put equal weight on the importance of the variables contributing to component one and component two. On the contrary, component one represents the largest amount of variation which can be explained by a linear combination of the variables and subsequent components explain less and less of the residual variation. Following the approach by Le Cao et al. (2008), the variables contributing to each component would be assessed separately. Variables contributing to the first component would be defined using the 'two stage average rank' method, afterwards the method would be applied separately to the second and any subsequent components. TThis approach was also used in section 8.4.5.

The diagram in Figure 9.1 describes the 'two stage average rank' method applied to the multiple Y variable and multiple component models. This method was based on the iterative process described by Eriksson et al. (2006a), Le Cao (2008) and González et al. (2011) adapted for the 'two stage average rank' method accommodating multiple Y variables and multiple components.

The data is split into a 'variable ordering training set' and 'variable selection training set'. Any missing data is imputed using NIPALS imputation. The initial step is to explore the number of components required by extracting 3, 4, 5, 10 and 100 variables and calculating the R^2-CV for one component. The R^2-CV for the addition of a second component is calculated for each Y variable using the variable ordering training set only. If the increase in R^2-CV is >=0.0975 for the addition of component two, then this indicates component two is contributing a significant amount to the prediction of that Y variable. The reason for the use of 0.0975 including a description of how to choose the required number of components is described in section 4.2.1.3.

Different numbers of variables (3, 4, 5, 10 and 100), were extracted in order to restrict the 1st component so it is not over fitted, allowing the second component to have more residual variation to explain by potential true predictors. This was explored in 8.4.5 but found not to affect the choice of number of components.

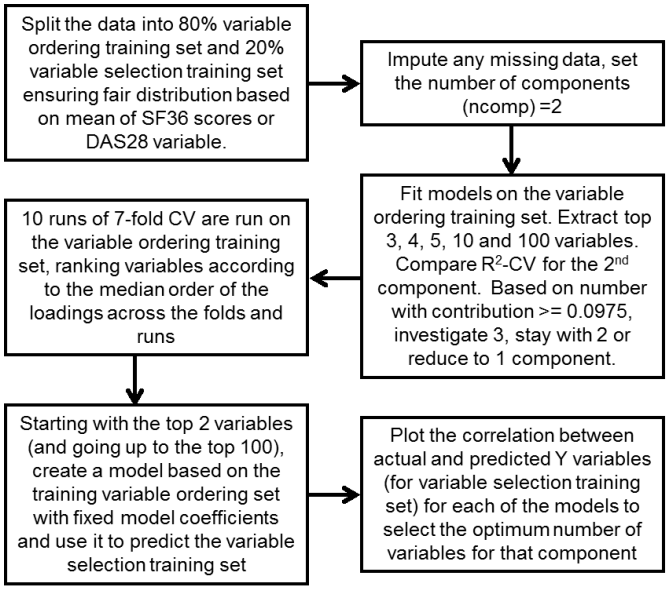


Figure 9.1 Multiple Y model fitting process using 'two stage average rank' method

As multiple Y variables are being modelled, the 2nd component may improve the model for some Y variables and not others. To define a consistent rule, it was decided that two or more components would be required if the additional R^2-CV explained by the 2nd component was >=0.0975 for more than one Y variable and consistently across restricting the 1st component to 3, 4 and 5 variables.

For the first component, the 'two stage average rank' method is used to select the number of variables to keep the first component. This consists of starting with the 80% 'variable ordering training set', fitting 10 runs of 7-fold CV and calculating the median rank for each variable from the 70 loading vectors corresponding to the 1st component. Using this median rank variable order, the first two variables are fitted in the model and the model coefficients estimated. This model is then used to predict the 'variable selection training set' (20% of the original dataset). The correlation between actual and predicted value for each of the Y variables is calculated. This method is then repeated extracting the top three variables, followed by the top four variables, up to 100 variables. For each Y variable, the number of X variables in the model is plotted against the correlation obtained from predicting the 'variable selection training set' to determine the optimum number of variables required. As the model is being tested on an independent set, it is protected against over fitting as observed in section 8.3.4, 8.4.5 and 8.4.8.

It is likely that the prediction of each Y variable may suggest a different optimum number of variables to include in the model. Therefore, the 'average' correlation across the Y variables is calculated. The average correlation is then used to determine the optimum number of variables which on average predicts the best for all of the Y variables together and this defines the final model.

Once the optimum number of variables for the 1st component is selected, if a 2nd component is required, then the 'two stage average rank' method is repeated restricting the 1st component to the number of variables suggested and using the same 'two stage average rank' method on the second component. The number of additional variables required for component two which optimises the correlation is selected as the final model. It was decided to let any variable selected in component

1 or component 2 be fitted in both components as if a variable is not important in a component, then it's coefficient will be close to 0 having little impact on the model prediction. If the 'three stage average rank' method was being used, the final model would be fitted on the 'variable selection training set', coefficients fixed and then used to predict an independent set of subjects.

9.3. Results of Y group 1: SF-36 analyses

The number of components was assessed and only one Y parameter (Mental Health) had an increase in R^2-CV associated with the 2nd component greater than 0.0975. The R^2-CV equalled just 0.1145 and was only observed when the 1st component was restricted to four variables for this one Y variable. All other SF-36 domains and extracting 3, 5, 10 and 100 variables for Mental Health all resulted in R^2-CV less than 0.0975 for the 2nd component. Therefore, this suggested that a second component wasn't beneficial and it was decided to remain with a one component model.

Models with from two to 100 variables were created fixing the coefficients on the 'variable ordering training set' and using the models to predict the 'variable selection training set'. The correlation between the predicted and actual SF-36 domains was calculated for each model and summarised using the average correlation across all domains. The maximum average correlation was observed using just the top three variables (Table 9.1 and Figure 9.2). The top three variables were; alcohol quantity used, alcohol use and symptom duration, however the average correlation between the actual and predicted Y variables was only r=0.159.

Investigating each of the SF-36 scores separately revealed that none of the domains are particularly well predicted on the variable selection training set (Table 9.1 and Figure 9.2). The highest correlation was achieved for Physical Functioning (r=0.330) using the top four variables of alcohol quantity, alcohol use, symptom duration and disease duration. However, the predictive ability of the model is poor as can be observed in Figure 9.3. It is of some interest that Physical Functioning was the best predicted domain given that severe RA would clearly interfere with physical activities such as bathing or dressing. However, this could be due to chance as Role Physical (r=0.161) is predicted poorly and it would be expected that the same physical deterioration would affect work and other daily activities as well.

Table 9.1 SF-36 models with optimum correlations based on which number of variables

Y Variable	Optimum correlation between actual and predicted values using test set	Number and name of variables to retain in the model
Average	0.159	(3) alcohol quantity, alcohol use, symptom duration,
Bodily Pain	0.141	(2) alcohol quantity, alcohol use,
General Health	0.144	(6) alcohol quantity, alcohol use, symptom duration, disease duration, age at onset of symptoms, BMI
Mental Health	0.183	(6) alcohol quantity, alcohol use, symptom duration, disease duration, age at onset of symptoms, BMI
Physical Functioning	0.330	(4) alcohol quantity, alcohol use, symptom duration, disease duration
Role Emotional	0.161	(25) Not listed
Role Physical	0.146	(19)
Social Functioning	0.121	(27)
Vitality	0.213	(4) alcohol quantity, alcohol use, symptom duration, disease duration

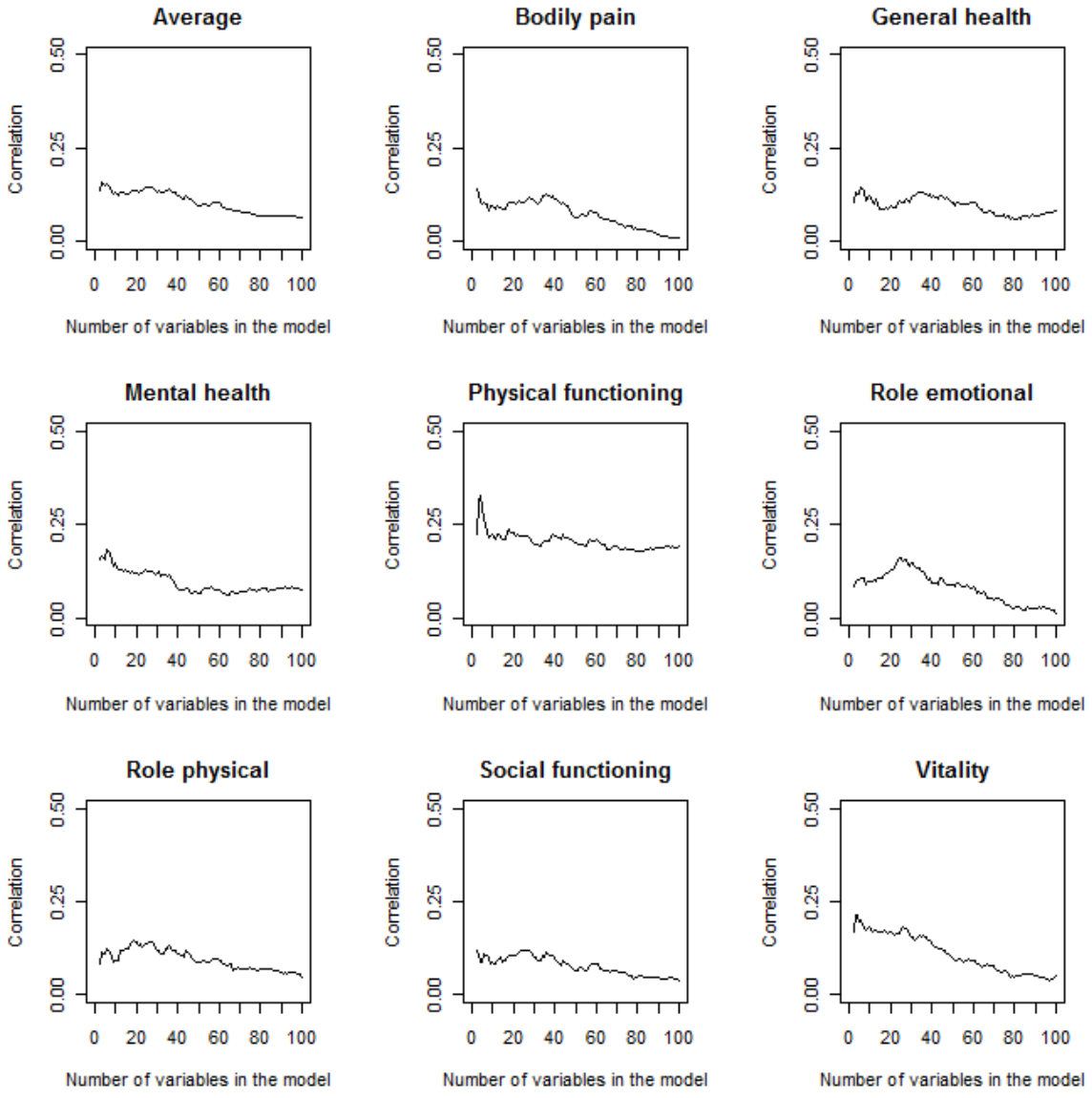


Figure 9.2 Correlation between actual and predicted SF36 domain for the test set

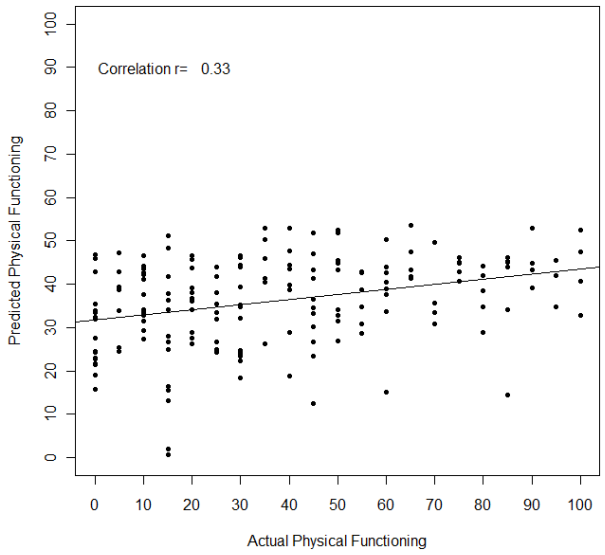


Figure 9.3 Scatter plot of actual versus predicted physical functioning for the test set

9.4. Results of Y group 2: SJC, TJC and DAS28 analyses

The number of components required was assessed and the increase in R^2-CV for each of the Y variables for all of the 3, 4, 5, 10, and 100 variables extracted was less than 0.0975. Therefore a one component model was deemed sufficient. None of the SJC, TJC or DAS28 variables were well predicted for the 'variable selection training set' when models were investigated using any number of variables and fixing the coefficients based on the 'variable ordering training set'. This can be observed from the very small Y axis scales in Figure 9.4 (range 0 to maximum of 0.09). On average, the following 11 variables (rs10785333, age, rs12035407, rs706778, rs2715038, BMI, rs11755527, rs8177374, rs12198924, disease duration, rs2073839) were suggested for the final model to obtain the best prediction, however this resulted in almost 0 correlation between the actual and predicted Y variables (r=0.064).

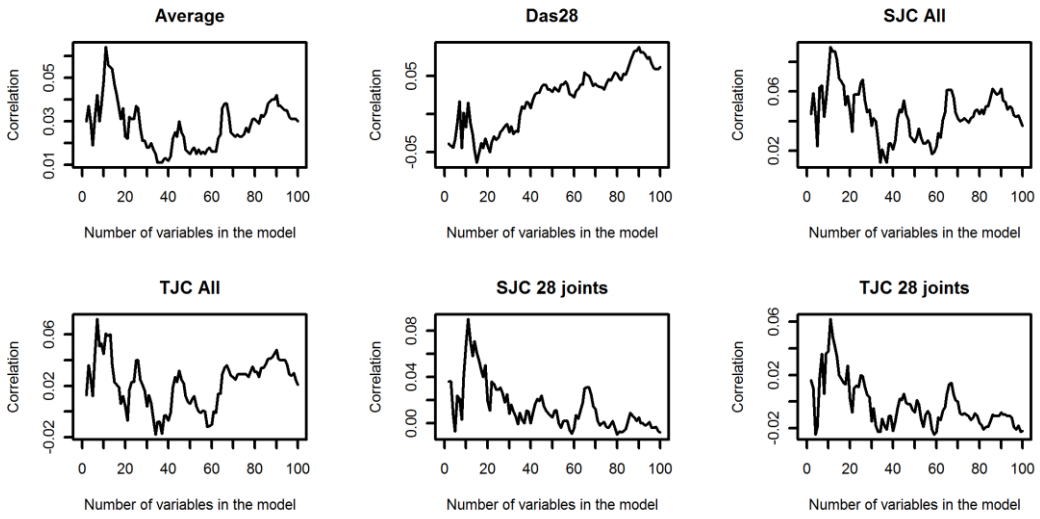


Figure 9.4 Correlation between actual and predicted TJC/SJC/DAS28 for the test set

9.5. Results of Y group 3: includes Larsen score, DAS28, MHAQ and others

This section describes the analyses of PVAS, RASEV, MHAQ, ESR, CRP, any erosions, the Larsen score (including the separate hand and foot counts) and DAS28. There was some evidence that two components would be beneficial in the model, particularly for ESR and CRP when the 1st component was limited to 5 variables or less. R^2-CV was greater than 0.0975 for ESR restricting the 1st component to 4 and 5 variables and CRP restricting the 1st component to 3, 4 or 5 variables. Therefore it was decided to investigate a two component model.

On average over the Y variables, the top 11 variables gave the best correlation of r=0.296 between actual and predicted Y variables for the variable selection training set (Figure 9.5). These variables corresponded to disease duration, symptom duration, age at onset of symptoms, age at time of diagnosis, ACPA category, ACPA value, alcohol quantity, alcohol use, BMI, rs26232 and rs2872507.

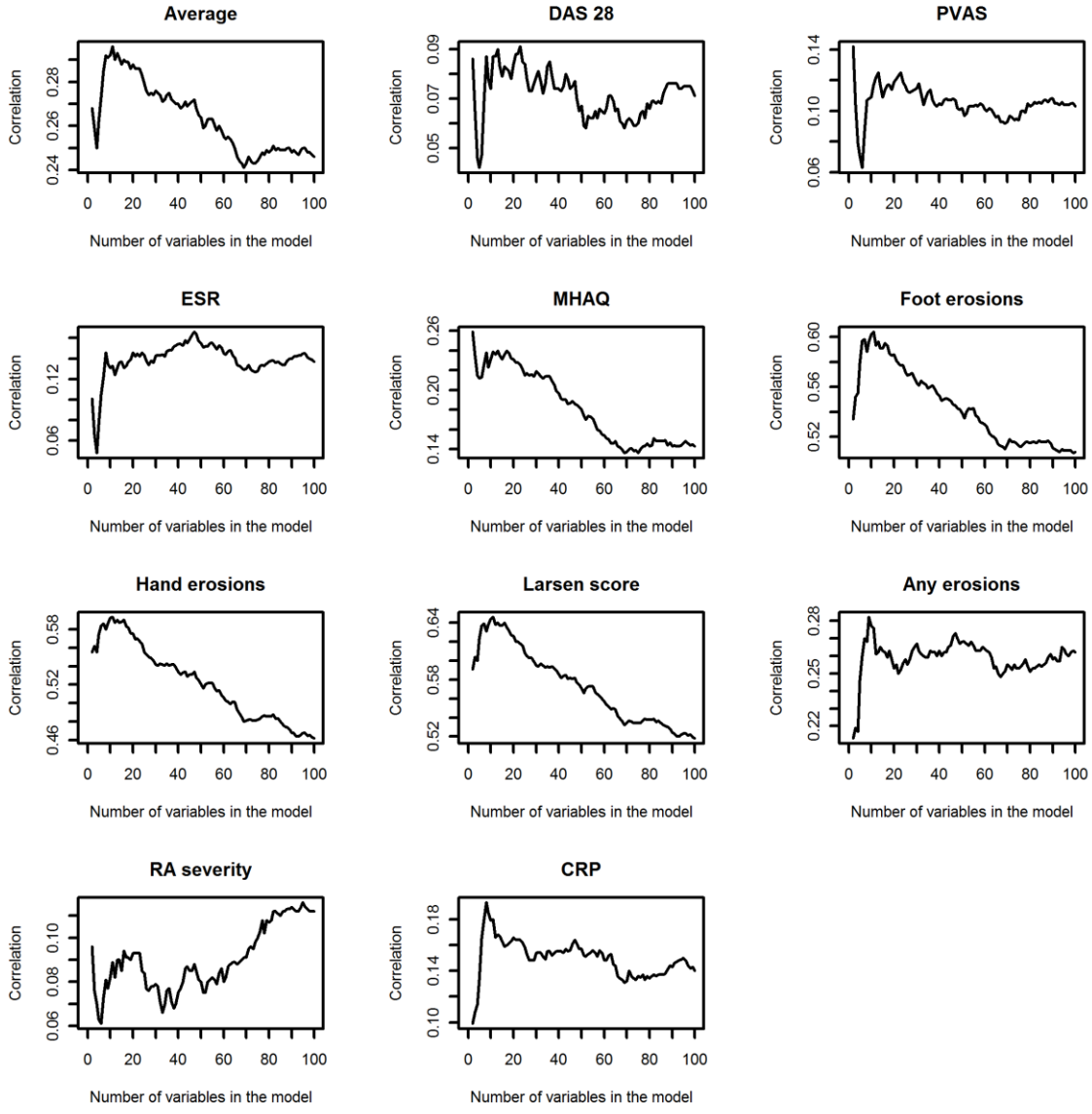


Figure 9.5 Correlation between actual and predicted Y variables for the test set

Noting the varying Y axis scales in (Figure 9.5), it was clear that the foot erosions, hand erosions & Larsen score variables were the best predicted, using 11 variables resulting in a maximum correlation of r=0.604, 0.593 and 0.646 respectively. DAS28 was particularly poorly predicted requiring 23 variables but only achieving a correlation of r=0.091. The remaining variables were also not well predicted as shown in Table 9.2.

Table 9.2 Optimum correlation models using one component based on number of variables

Y Variable	Correlation[a]	Number and name of variables to retain in the model
Average	0.296	(11) disease duration, symptom duration, age at onset of symptoms, age at time of diagnosis, ACPA category, ACPA value, alcohol quantity, alcohol use, BMI, rs26232 and rs2872507
DAS28	0.091	(23) not listed
PVAS	0.142	(2) disease duration, symptom duration
ESR	0.166	(47) not listed
MHAQ	0.259	(2) disease duration, symptom duration
Foot erosions	0.604	(11) disease duration, symptom duration, age at onset of symptoms, age at time of diagnosis, ACPA category, ACPA value, alcohol quantity, alcohol use, BMI, rs26232 and rs2872507
Hand erosions	0.593	
Larsen score	0.646	
Any erosions	0.282	(9) disease duration, symptom duration, age at onset of symptoms, age at time of diagnosis, ACPA category, ACPA value, alcohol quantity, alcohol use, BMI,
RA severity	0.116	(95) not listed
CRP	0.193	(8) disease duration, symptom duration, age at onset of symptoms, age at time of diagnosis, ACPA category, ACPA value, alcohol quantity, alcohol use,

a: Table shows the N variables required to achieve the optimum correlation between actual and predicted Y variables calculated using the test set with one component.

Based on the best average correlation, the top 11 variables were selected to be retained for the 1st component. These were fitted along with all variables for component two and the average ranked loading vectors corresponding to the 2nd component investigated from the 10 runs of 7-fold CV. The maximum correlation was observed using 16 variables in total, increasing the average correlation from r=0.296 (with 11 variables and one component) to r=0.353 (with 16 variables and two components) (Table 9.3). The top 16 variables were; disease duration, symptom duration, age at onset of symptoms, age at time of diagnosis, ACPA category, ACPA value, alcohol quantity, alcohol use, BMI, rs26232, rs2872507, rs2715038, rs11755527, age, rs4892117 and rs3218253. The first 11 in this list correspond to those selected in the one component analysis but all variables are allowed to contribute to both components.

There was very little increase for the foot erosions, hand erosions & Larsen score variables with the best correlation with two components obtained with 15 (r=0.608), 21 (r=0.604) and 18 (r=0.653) variables (compared with one component with 11 variables r= 0.604, 0.593 and 0.646 respectively. This very slight improvement with the addition of many more variables and two components confirms that for Larsen score modelling, only one component is required (as observed in chapters 4 and 6). However, the inclusion of a 2nd component did increase the correlation for most of the other variables including CRP and ESR. Both of these variables had an R^2-CV for the 2nd component greater than 0.0975 indicating the 2nd component was required for prediction of these variables. The largest increase in correlation corresponded to the DAS28 variable (Table 9.3). In general, the

two component model gave better predictions on average across the Y variables but only by an increase in correlation of 0.057 when predicting the 'variable selection training set'.

Table 9.3 Models with optimum correlations for one or two components

Y Variable	N variables: correlation one component[a]	N variables: correlation two components[a]
Average	11: 0.296	16: 0.353
DAS28	23: 0.091	16: 0.229
PVAS	2: 0.142	15: 0.250
ESR	47: 0.166	17: 0.262
MHAQ	2: 0.259	14: 0.289
Foot erosions	11: 0.604	15: 0.608
Hand erosions	11: 0.593	21: 0.604
Larsen score	11: 0.646	18: 0.653
Any erosions	9: 0.282	102: 0.295
RA severity	95: 0.116	36: 0.198
CRP	8: 0.193	11: 0.224

a: Table shows the optimum N variables required to achieve the optimum correlation between actual and predicted Y variables calculated using the test set with one or two components.

Although the 'variable selection training set' has already been used to select the number of variables which will form the final model (per the 'two stage average rank' method) it was decided using the chosen 16 variables, to fit a PLS model on this set and estimate the model coefficients. This final model was then used to predict the 'variable selection training set', despite knowing this is likely to be an overestimate of the model prediction ability if applied to an independent set. The 183 test subjects predicted values are shown in Figure 9.6.

The multiple Y variable model revealed a better correlation (r=0.651) between predicted and actual Larsen score compared to modelling the Larsen score variable alone in section 8.4.3, where the maximum correlation obtained was r=0.576 with one component. It is worth noting that the split of patients into a 'variable ordering training set' and 'variable selection training set' in section 8.4 was based on the Larsen score distribution and the split in this section was based on DAS28 distribution. Hence, the difference in prediction ability could be due to the split of patients, the additional benefit of modelling multiple variables together or just random variation. In order to investigate further, this section was reproduced using the 'variable ordering training set' and 'variable selection training set' based on the Larsen score as used in section 8.4.

Using the 'two stage average rank' method, the optimum prediction results were found with a one component model containing the top eight variables; disease duration, symptom duration, age at onset of symptoms, age at time of diagnosis, ACPA category, ACPA value, alcohol quantity and alcohol use. This obtained a correlation between the variable selection training set actual and predicted Larsen score of r=0.583. Investigation of a 2nd component found the best average prediction across all of the Y variables to be based on the top 67 variables, however this decreased the prediction accuracy of the Larsen score to r=0.564. Hence, both models agree that for the Larsen score, only one component is required although a second component is beneficial for other Y variables. The variables selected by using the different 80:20 splits of data are very similar for the 1st component (top 8 variables are identical). Therefore it appears to be random variation

associated with which patients happen to be in the 20% predicted which impact on the selection of lower order variables which are only contributing a small amount to the model.

These results suggest that there is no evidence that modelling multiple Y variables together improves the ability to predict the Larsen score. There is some variation in prediction ability depending on how the 'variable ordering training set' and 'variable selection training set' are created. Ensuring that the Larsen score is balanced, resulted in a correlation of r=0.583 using an 11 variable model and ensuring that DAS28 is balanced, resulted in a correlation of r=0.651 using an 8 variable model (8 variables are the same). An area of future research would be to quantify the variation around the correlation achieved using different splits of the data as described by Daetwyler et al. (2013). These methods are discussed in the areas for future research in section 10.4.

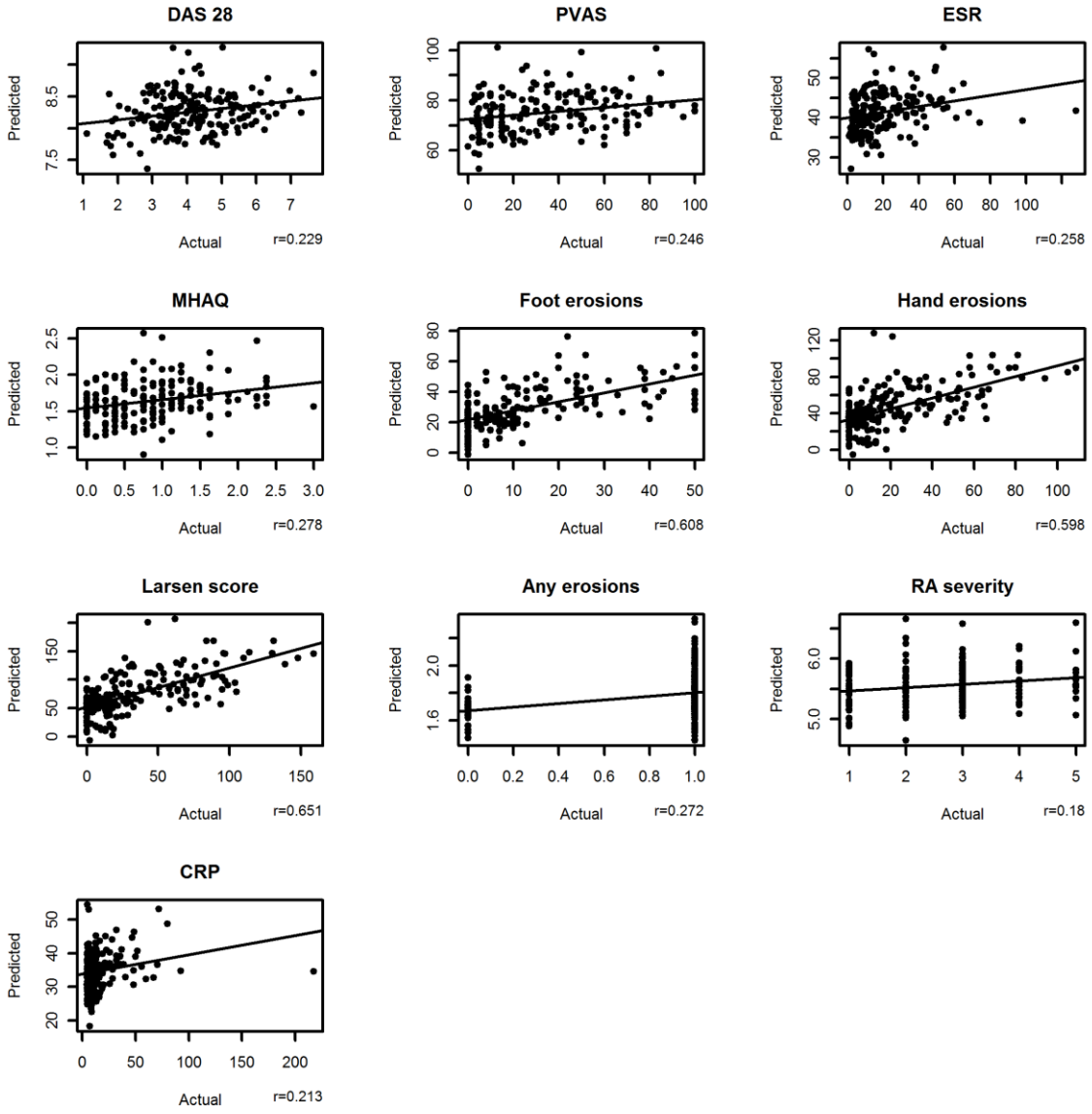


Figure 9.6 Actual versus predicted Y variables based on two components and 16 variables

In order to investigate the stability of the models created, it was decided to fit the 'three stage average rank' method using both the Larsen score and DAS28 score to split patients into the 40%:40%:20% sets. Although modelling multiple Y variables, primary focus would be on prediction of the Larsen score, as prevous research revealed poor prediction of the other severity variables. Only one component was explored as the second component was found not to add adequate Larsen score predictive ability in section 8.4.6 or earlier in this section.

Table 9.4 reveals that the top seven variables selected using the 'three stage average rank' method are the same across models formed using DAS28 or the Larsen score to determine the split of the data. In addition, the top seven variables using a 80:20 split or a 40:40:20 split are also consistent. As may be expected, variables which have a smaller effect are more variable when looking at different splits of the data as more patients are required to detect smaller effects than larger effects. Table 9.4 could suggest that 40% of the data to initially rank variables may be too small, as the two stage model split based on DAS28 achieved a better Larsen score correlation than the three stage model split by DAS28. However, as the same was not observed by splitting the data by the Larsen score and using a two or three stage model, it is believed to be simple variation observed by using a different 20% of patients to predict. Note that the 20% in the two stage model was not the same 20% in the three stage model.

Table 9.4 Comparison of using different splits of the data for multiple Y variable modelling

Model	N Vars	Variables selected in 1st component	Larsen score correlation
2 stage: 80:20 split based on DAS28	11	Disease duration, symptom duration, age at onset of symptoms, age at time of diagnosis, ACPA category, ACPA value, alcohol quantity, alcohol use, BMI, rs26232, rs2872507	0.646 [1]
2 stage: 80:20 split based on Larsen score	8	Disease duration, symptom duration, age at onset of symptoms, age at time of diagnosis, ACPA category, ACPA value, alcohol quantity and alcohol use	0.583[1]
3 stage: 40:40:20 split based on DAS28	10	Disease duration, symptom duration, age at onset of symptoms, age at time of diagnosis, ACPA category, ACPA value, alcohol quantity, alcohol use, BMI, Average number of cigarettes per day.	0.547[2]
3 stage: 40:40:20 split based on Larsen score	13	Disease duration, symptom duration, age at onset of symptoms, age at time of diagnosis, ACPA category, ACPA value, alcohol quantity, DRB1 S1, rs2872507, rs9366826 rs7234029 rs394581 rs3788013	0.584[2]

[1] indicates correlation calculated using 20% subjects who have already been used to determine the number of variables to retain in the model.
[2] indicates correlation calculated using 20% subjects who are independent from all model creation.

9.6. Summary

At the start of this research (section 2.2.1.2), it was speculated that the Larsen score would be the best measure of severity as it is more stable over time than other laboratory or quality of life measures. This chapter modelled multiple Y variables together with the anticipation of improved predictive performance. However, the models revealed very poor prediction of all of the SF-36 domains, SJC, TJC, DAS28, PVAS, ESR, MHAQ, any erosions, RA severity (RASEV) and CRP. In fact, the Larsen score (including when it was broken down into hand and foot erosion counts), were the only severity variables which obtained a correlation over 0.5 between the actual and predicted values for the 'variable selection training set'.

PVAS, RASEV, MHAQ, ESR, CRP, any erosions, the Larsen score (including the separate hand and foot counts) & DAS28 were modelled using the 'two stage average rank' method described in section 8.3.3. The method was amended to accommodate multiple Y variables as described in section 9.2. A two component model was deemed necessary. Eleven variables in the model resulted in the highest average correlation across all of the Y variables using just the first component (r=0.296). These 11 variables consisted of disease duration, symptom duration, age at onset of symptoms, age at time of diagnosis, ACPA category, ACPA value, alcohol quantity, alcohol use, BMI, rs26232 and rs2872507. Addition of a second component suggested five further variables were required to improve the average correlation. Therefore the model consisted of 16 variables in total contributing to the 1st and 2nd components. The average correlation increased by 0.057 (to r=0.353) with two components and the following additional variables; rs2715038, rs11755527, age, rs4892117 and rs3218253. Whilst the 2nd component did not contribute much to the prediction ability of the Larsen score (or separate hand and foot counts), it did enable improved prediction of some of the other severity measures (particularly DAS28 which had a correlation of r=0.091 with one component and r=0.229 with two components). This supports the previous work modelling the single Larsen score variable where only one component was required.

It was noted that the prediction of the Larsen score improved using multiple Y variables in the model compared to the single Larsen score. However, after investigation using both DAS28 and the Larsen score to split the data and investigating using the 'two stage average rank' method and the 'three stage average rank' method, differences in the correlation were presumed due to using a different split of the data and not the method used. This highlighted the sensitivity of correlation estimation when different subsets of data are used and an area of future analysis would be to explore the variation around the correlation estimate. Therefore, there was no evidence to suggest that the Larsen score prediction was improved by modelling multiple Y variables together.

It is speculated that the inability to use this data to accurately predict the multiple measures of RA severity could be due to the cohorts varied disease duration. Measures of disease activity (such as CRP and ESR) may not reflect the severity over time. Patients who are susceptible to a more severe disease but are currently controlled due to treatment, may have the effect of their genetic variants too confounded with the treatments they have received over a long period of time. It would therefore be beneficial to repeat this work using a large cohort of subjects who have been followed up from disease diagnosis over a period of time so that the rate of their progression can be measured. As treatment regimens over this period would be recorded, these could be adjusted for in the model and it is possible that this would enable a more sensitive detection of genetic variants affecting RA severity.

10. Conclusions

10.1. Summary of the motivation behind the research

The aim of this project was to attempt to create a model, which could be used at RA disease diagnosis, to predict whether a subject is at high or low risk of developing the more severe form of RA. The project was driven by the availability of a large dataset measuring severity of RA, genetics and environmental factors at a cross-section in time on 1009 subjects. Patients had varied disease duration from one to 65 years and varied severity of disease from zero to 160 as measured by the Larsen score. The Larsen score is a radiographic measure of cartilage and bone damage. There were a high proportion of patients with zero Larsen score (13.6%) and patients appeared to have developed RA at a younger age than expected in the population particularly for males (mean of 47 years old at disease onset). Despite this, the cohort appeared representative of the population in terms of gender (73% female), BMI (mean=26), smokers (20% smokers and 39% former smokers) and alcohol use (68% drinkers). The mean age of the cohort was 61 years with the youngest patient aged 20 and eldest 92 years old.

Data from a GWAS contained 325,482 SNPs and was available on 394 subjects. For this reason the data was split into two sets. The first set consisted of 912 subjects with available Larsen score data, 19 environmental measures and 368 SNPs and the second set consisted of 394 subjects, 19 environmental measures and 325,482 SNPs. After exploration of multivariate modelling methods, PLS was selected as the most appropriate methodology for this research for the reasons summarised in Section 3 and 10.3. Modelling methods were developed based on the smaller dataset ('all subjects' dataset) and then applied to the larger set ('GWAS SNPs' dataset). It was decided to initially focus analysis on the Larsen score as the primary measure of RA severity, as erosive damage is more stable over time than measures of inflammation. The analysis was later extended to multiple Y variables.

10.2. Summary of results

Many authors select variables in PLS models using a type of 'percentage fold' method often insisting variables have to be selected in 100% of the folds and runs (González et al., 2011, Le Cao et al., 2008, Eriksson et al., 2006a, SAS, 2008, Long et al., 2011). Multiple runs are used to avoid model selection being based on a particular fold on the data. Whilst this appears to work well for variables with big effects sizes, this research found this to be ineffective for SNP data where smaller effects were being modelled and rare variants often led to a SNP not even being included in some folds due to insufficient variation (Figure 6.7). The optimum number of variables to select for a model was often difficult to determine especially to avoid over fitting. Using an 'average rank' method (simply ordering the variables based on the average rank of their loading) was more sensitive and allowed more SNPs to be considered for entry to the final model, despite not being selected by all folds or runs of the data.

It is also common to investigate the number of variables to retain in the final model by plotting the R^2-CV by the number of variables included in the model and selecting the optimum as the number of variables to extract (González et al., 2011). This appears to work well for smaller datasets, for example, detecting a few key predictors with large effects from less than 400 variables. However, as SNPs had to be modelled in groups of less than 12000 in R, this was very time consuming as each graph had to be inspected manually. The 'average rank' method removed this requirement by

simply selecting any variables on average ranked less than 200 and carrying these variables through to a higher level model. Using this method, CV was shown to not protect against over fitting of the model (Figure 6.8). Hence, too many variables were included in the final model which resulted in an overestimate of the predictive ability and a model which would not transfer well to an independent set. This was apparent when the model was re-created on 80% of the data and tested independently on the 20% left out (Figure 8.6).

Random permutations of the data showed that there was evidence that the model including environmental and SNPs performed better than chance alone and in particular that the SNPs were also contributing to the prediction. Therefore, methods were explored to try to avoid creating a PLS model which suffers from over fitting.

Initially, a 'two stage average rank' method was used which split the 'GWAS SNPs' dataset into 80% for deriving the order of importance of the variables and 20% for selecting the optimum number required for the final model. The data was split using the distribution of the Larsen score to ensure fair representation of the spread of the data in both sets. Unfortunately, this method suggested the best model would be one containing disease duration, symptom duration and age at time of diagnosis only. Therefore, it appears that PLS is unable to select any SNPs from the 'GWAS SNPs' dataset for inclusion in a prediction model. It is believed this is partly due to the sample size being used (N= 394 subjects) and partly due to wide unexplainable variation in the data. The GoRA data used come from a cross-sectional cohort with very varied disease durations and different treatment regimens received. Many variables which may account for these differences amongst subjects are not available for analyses and hence variation in the Larsen score is unexplainable. The same PLS approach would be recommended to be applied to a more homogenous cohort. It is hoped that even with a sample size around 400, that key SNP effects would be able to be identified.

It was decided to apply the same methods developed on the 'GWAS SNPs' dataset to the 'All subjects' dataset, to investigate if SNPs could be identified as predictive of RA severity when assessed using a larger sample size.

A 'two stage average rank' method was able to control against over fitting the data and resulted in a model containing seven variables which were well documented predictors of RA severity (disease duration, symptom duration, age at onset of symptoms, age at time of diagnosis, ACPA category, BMI and ACPA value). However, no SNPs were selected for the final model. Estimation of the correlation between actual and predicted Larsen score was r=0.576, which is likely to overestimate the true prediction ability, as the 20% independent set were used to select the optimum number of variables for the final model. Of interest was that the model was not able to predict a Larsen score value greater than 85, suggesting that the key predictors of extremely high severity were not included in the model. The ability to predict the extremes of the distribution of the Larsen score would need to be investigated in another cohort, to truly determine whether this is a consequence of PLS not predicting non normally distributed data very well, or due to missing key predictors in the model and wide unaccountable variation.

To investigate how the model performs on a truly independent set of data, a 'three stage average rank' method was developed which split the data into a 40% 'variable ordering training set', a 40% 'variable selection training set' and a 20% independent test set. Although a slightly different model was obtained containing 10 variables (disease duration, symptom duration, age at symptom onset,

Age at diagnosis, ACPA category, ACPA value, rs26510, BMI, DRB1 S2 and rs26232), similar correlation between actual and predicted Larsen score was found (r=0.559). Of particular interest was that three genetic variants were selected to be in the model and all except rs26510 (ERAP-1) are found in the literature predictive of RA severity. This suggested that with 912 subjects the 'three stage average rank' method is able to identify genetic variants contributing to RA severity. If the sample size was increased and more SNPs were recorded on a more homogenous cohort, it would be expected more variants could be identified using this method.

Unfortunately, attempting to use this dataset on subsets of patients, to attempt to homogenise the cohort and reduce some of the unexplainable variation, did not lead to much success. Investigation of subjects with a disease duration of <10 years and separately of <15 years, resulted in a lower correlation between actual and predicted Larsen score than was achieved modelling the full cohort of patients together. Investigation of the ACPA positive subset (N=689) did increase the ability to predict the independent 20% set of data (r=0.611), however the model only included the following six variables (disease duration, symptom duration, age at onset of symptoms, age at time of diagnosis, BMI and rs2073839). Using the 'three stage average rank method' meant that just 40% (N=275) are used to create an order of importance for the variables and select the optimum number of variables. It was believed that this may not be sufficient for SNPs with smaller sized effects to be identified. Hence, the positive effect of reducing the heterogeneity may be overshadowed by also reducing the sample size.

A particular advantage of using PLS methods is its ability to model multiple Y variables together. However, attempts to model multiple severity measures on the GoRA cohort had limited success. One reason for this could be because severity measures such as the Larsen score tend to be less variable over time than laboratory measures (such as ESR and CRP) which fluctuate with disease activity. Therefore, it would be interesting to explore this further in patients with similar disease duration.

The variables which were selected as predictors of RA severity for each of the models summarised in Table 10.1.

Using the 'GWAS SNPs' dataset, the best prediction of the test set was obtained using just disease duration, symptom duration and age at time of diagnosis. These are all well-known predictors of severity. Modelling the 'All subjects' data using a 'three stage average rank' method extended this list to include age at onset of symptoms, ACPA category, ACPA value, BMI, DRB1 S2, rs26232 and rs26510. All of these variables (except rs26510) are documented in the literature review to be predictive of RA severity (section 2.3 and 2.4). Although the <10 years and <15 years disease duration subgroup analyses were unsuccessful, the ACPA positive model achieved slightly better correlation than the 'all subjects' dataset analysis (r=0.611). The model contained disease duration, symptom duration, age at disease onset, age at diagnosis, BMI and rs2073839. rs2073839 is found in the intronic region of the solute carrier family 22 (SLC22) A4 gene in chromosome 5 and all other variables are known predictors of RA severity identified in the literature review.

Table 10.1 Summary of final models

Y variable(s)	Data: model	Method & Variables[b]	Correlation[c]
Larsen score	'GWAS SNPs' 394 subjects	2 stage 80/20: DD, SD, Agediag	Larsen=0.622
	'All subjects' 912 subjects	2 stage 80/20: DD, SD, Ageonset, Agediag, ACPA category, ACPA value, BMI	Larsen=0.576
		3 stage 40/40/20: DD, SD, Ageonset, Agediag, ACPA category, ACPA value, BMI, rs26510, DRB1 S2, rs26232	Larsen=0.559
	Disease duration <10 years, 350 subjects	3 stage 40/40/20: DD, BMI, rs443198, rs2568127, rs4133002, rs4535211	Larsen=0.284
	Disease duration <15 years, 535 subjects	3 stage 40/40/20: DD, ACPA category	Larsen=0.424
	ACPA positive, 689 subjects	3 stage 40/40/20: DD, SD, Ageonset, Agediag, BMI, rs2073839	Larsen=0.611
Multiple Y[a]	'All subjects' split using DAS28: one components, 914 subjects	2 stage 80/20: DD, SD, Ageonset, Agediag, ACPA category, ACPA value, alcohol quantity, alcohol use, BMI, rs26232, rs2872507	Average=0.296 Larsen=0.646
	'All subjects' split using DAS28: two components, 914 subjects	2 stage 80/20: DD, SD, Ageonset, Agediag, ACPA category, ACPA value, alcohol quantity, alcohol use, BMI, rs26232, rs2872507, rs2715038, rs11755527, age, rs4892117, rs3218253	Average=0.353 Larsen=0.651

[a] PVAS, RASEV, MHAQ, ESR, CRP, any erosions, Larsen score (hand and foot counts) & DAS28
[b] Method implies whether a 'two stage average rank' method using 80% & 20% sets was used or a 'three stage average rank' method using a 40%:40%:20% split of data. DD=disease duration, SD=symptom duration, Agediag=age at time of diagnosis, Ageonset= age at onset of symptoms.
[c] Either correlation between actual and predicted Larsen score or average correlation between actual and predicted results for all Y variables included in the model.

Although modelling the multiple Y variables did not reveal any improvement in the correlation between actual and predicted Larsen score, some SNPs were selected as contributing to the final model. These were; rs26232 (see C5orf30 in section 2.3.2.6), rs2872507 IKZF3 gene RA risk allele (Stahl et al., 2010, Kurreeman et al., 2012), rs11755527 BACH2 gene Type I diabetes risk allele (Cooper et al., 2008), rs4892117 investigated for a RA susceptibility gene (Barton et al., 2008) and rs3218253 IL2RB gene RA risk allele (Stahl et al., 2010).

In conclusion, a 'three stage average rank' method appears to be able to reduce large numbers of environmental variables and SNPs to select the most predictive of the Larsen score. However, its success is dependent on the quantifiable variability in the data, the sample size available for analysis and the size of the SNP effect. A discussion of why PLS is more appropriate in this context of research than using more traditional regression based method is discussed in Section 10.3. Section 10.4 provides further details of how this work has contributed to the research area and section 10.5 details areas of future research including a consideration of other methods which could be used.

10.3. Justification of the choice of PLS methodology

Although PLS was chosen as the primary method for analysis at the start of the research (section 3), other methods could have also been investigated such as random forests, structural equation modelling or other penalised regression methods such as LASSO or Elastic nets. These are described as areas for future research in section 10.5, as given the size of the dataset being explored, there was insufficient time to explore these methods in this research.

The 'All subjects' dataset had more observations (N=912) than variables (387). This indicates that it could be modelled using a more simple regression based approach. However, even on a relatively small dataset, numerous problems were encountered which makes the methods inappropriate to apply to the 'GWAS SNPs' dataset. This supports the use of the more complex PLS approach. The key issues are summarised below.

Instability of p-values: It is well documented that when analysing correlated variables in a linear regression model, the estimates for the model parameters become unstable (Geladi and Kowalski, 1986). However, problems with instability also arose when analysing the SNPs one at a time in a model containing environmental variables to try to predict the single Y variable of the Larsen score. Using the 'All subjects' dataset (N=912), a ZINB regression model appeared to be appropriate to model the inflation in the number of subjects with zero Larsen score. However, the p-values became unstable, when all subjects with a Larsen score of 0 had a genotype of 0 (most frequent homozygous). For example, rs2071592 was highly significant in one analysis using NIPALS imputation (section 4.4.2, $p=9.2 \times 10^{-11}$) and then not significant when using Quick imputation (section 4.3.2, p=0.453). The only difference between these models was the way in which 142 subjects with missing data were imputed. Whilst the quick imputation assigned all subjects to have a zero and thus retained the problem that all subjects with a Larsen score of 0 had a genotype of 0, the NIPALS imputation assigned them values between 0 and 2. The PLS analysis was not sensitive to the method of imputation retaining the variable in both models, however, a very different conclusion is formed using univariate p-values to indicate the importance of a variable.

Despite this intial issue, two further methods of regression analyses were explored. A forward selection ZINB regression strategy for the 'All subjects' dataset (discussed below) and using univariate NB regression models for the 'GWAS subjects' dataset (section 6.5.5).

Time constraints: Whilst forward or stepwise ZINB regression selection methods are easily implemented in R, it requires all variables to be fitted in the model before using the stepAIC function. This is not possible for this data as there are too many variables to fit in one model and the model does not converge. Therefore, to perform stepwise regression, a macro was written to take each variable in turn, fit the variable into the null model and evaluate its significance. However, running the 387 variables for the 'All subjects' dataset consisted of fitting 387 ZINB regression models which took 11 minutes. For this reason, the forward stepwise regression technique would not be practical when analysing the 325,482 'GWAS SNPs' dataset. For example, it would take an estimated 154 hours (6.4 days) to run just the first stage of selecting the first variable for the GWAS data. Should the final model contain 30 variables, this would take 4620 hours or 192.5 days and this does not include exploration into removing the variables from the model if they become non-significant. This is potentially why pre-filtering of SNPs is often used with penalised regression techniques to remove correlation and limit the number of variables required to be

investigated. However, as PLS using SIMCA can be run in one hour on the entire 'GWAS SNP' data and this provides an ordered list of importance of the variables, it is considerably more efficient than regression methods and allows all potential variables of interest to be investigated. Table 10.2 provides a summary of the running times for each of the PLS models analysed in R. Whilst the average rank method (using 1 stage) takes just 8 hours, applying the two or three stages takes only five to ten minutes longer. This is due to the number of variables included in the models for the 2nd and 3rd stages is substantially reduced from the thousands investigated in the first stage.

Table 10.2 Running times for various PLS models

Model fitted	'GWAS SNPs' data 325482 SNPs, 394 subjects Run in 40 separate blocks	'All subjects' data 368 SNPs on 912 subjects Run in 1 block
Percentage fold method: 50 runs of 10 fold CV	Minimum 1 week Each of the 40 blocks, takes approximately 55 minutes to determine variables required to extract + manual time to look at graph, followed by 70 minutes to run 50 runs of 10 fold CV extracting the appropriate number of variables and creating final model. 125 mins per block=5000 mins (83 hrs) however as this cannot be run over night due to requiring manual intervention. It takes a minimum of a week to run.	Minimum 30 minutes 13 minutes to determine variables required to extract + manual time to look at graph 13 minutes to run 50 runs of 10 fold CV extracting the appropriate number of variables.
Average rank method: 50 runs of 5 (or 10) fold CV	50 hours continuous	20 minutes continuous
1 stage: 10 runs of 5 (or 10) fold CV	8 hours continuous on PC (5 hours on a Linux cluster machine)	5 minutes

Selection of variables: Even if it was possible to program a forward stepwise ZINB regression model more efficiently and thus reduce the time it takes to apply it to the 'GWAS SNP' dataset, modelling correlated variables introduces additional problems. The most significant variable is entered into the model first, after which other largely correlated variables may become none significant. Therefore potentially important variables are not retained for the final model, even though they may be able to explain some of the severity. In addition, it is not known whether the variable selected for the model is a surrogate for another important variable. For example, when modelling the 'All subjects' dataset using forward selection, disease duration and symptom duration were added to the ZINB model (as the 1st and 9th variables entered), however, age at symptom onset and age at diagnosis were not entered at all. All four of these variables (known predictors of RA severity) are included in PLS models because it is capable of retaining correlated variables in the same model thus increasing their ability to contribute to prediction.

Number of variables in a single model: Linear regression models are limited by the sample size available to contribute to the number of degrees of freedom which can be 'spent' on the variables fitted in the regression model. A ZINB regression model using the 'all subjects' dataset (N=912), was only able to model the top 30 variables (in both the count and zero inflation part of the model) before having convergence issues. Potentially, there are many SNPs having a small but important effect on phenotypes. PLS has the advantage of using a dimension reduction technique which

models the components rather than the variables directly. This allows all variables to be investigated in the same model.

Univariate SNP testing: Instead of selecting SNPs using a forward or stepwise selection, section 6.5.5 investigated using univariate NB regression models for the 'GWAS subjects' calculating the p-value corresponding to fitting each SNP on its own in a model containing some key environmental variables. Unfortunately, with just 394 subjects, there was not enough power after adjusting for multiple testing of 325482 SNPs. Hence, no SNPs were found with a p-value less than the bonferoni adjusted significance level. Although the top 100 SNPs corresponding to the lowest p-values could be used to form a prediction model and this was shown to perform as well as the top 100 SNPs selected by PLS, there is still no method to prevent over fitting or determine the optimum number of variables to retain. In addition, each SNP tested univariately is selected due to its perceived influence on severity. Correlated variables may be explaining the same variability and hence if these are then used to form a prediction model (i.e. the SNPs entered into the same model), they may be some overlap resulting in collinearity issues. PLS does not suffer from these collinearity problems (due to modelling the loadings) and it selects variables to explain as much as the variability as possible. It is therefore likely to create better predictive models which explain more of the variation in the Y data.

This research provides evidence that PLS is a quick model fitting strategy for creation of a prediction model, when data contain many more variables than observations. PLS is not limited by standard regression model assumptions, this allows many correlated variables to contribute to the model.

10.4. Contribution of this work to current research and limitations

Genotyping costs have dramatically reduced over the last 10 years and this has led to the amount of genetic data for analysis increasing substantially. As this is a relatively new phenomenon, methods to deal with the data are continually being developed. A basic review of the literature using the search terms 'PLS' and 'SNP' did not find any papers prior to 2010 investigating more than 20,000 SNPs in the same analysis. No studies were found using PLS models to predict RA severity using environmental factors and genetic variants.

Since 2010 much of the research has been in animal studies (such as breeding or milk yield in cattle) or in plant studies and the number of SNPs being investigated together is ever increasing. Recent modelling strategies have addressed specific issues, however none have reported a method within PLS models to avoid over fitting and enable modelling of high-dimensional data without pre-filtering. Recently, there have been warnings against the use of pre-filtering as it could lose important signals in the data (Abraham et al., 2013).

Daetwyler et al. (2013) reviewed the current methods being applied in plant and animal genomic prediction. They discuss two forms of CV, that of separate training and test sets (either run once or replicated) and that of replicated internal CV (such as the n runs of k-fold CV). They highlight that correlation between the estimated and true response is commonly being employed to quantify prediction accuracy and state that CV appears to be the preferred validation method. They warn that CV only gives an estimate of the accuracy of the model in the data it is applied to, hence may not be transferable to other sets of data. The use of averaging of CV replications to avoid bias

caused by the choice of folds has also been suggested for penalised regression methods and this approach has been applied to the PLS models in this PhD research (Li and Sillanpaa, 2012).

Current methods developed in the SPLS area (Le Cao et al., 2008, Magidson, 2011, González et al., 2011) use internal CV and specify the maximum number of variables to retain in the final model. Variables are selected for the final model if they are chosen in a certain number of folds across all of the runs entitled the 'percentage fold' method. However, this research demonstrated that this method does not protect against over fitting and it is very time consuming when applied to a high-dimensional GWAS dataset.

Long et al. (2011) analyse 32,518 SNPs which they use to predict milk yield in Holstein cattle. They use the SPLS tuning parameters (Kappa and Eta) to identify the models with the best correlation under CV however, because the tuning parameters can lead to models with a large number of SNPs, they suggest specifying an upper bound for the number of SNPs required in the final model. This is similar to the recommendations using mixOmics and COREXpress (Le Cao et al., 2008, Magidson, 2011, González et al., 2011). Following the upper bound specification, they use a range of tuning parameter values to investigate models which have less than this maximum number of SNPs. They comment that investigating all of the combinations of (Kappa and Eta) may be computationally expensive, hence the requirement to pre-specify the maximum number of variables required to be retained for the final model. In addition, they have substantially fewer SNPs than are being investigated in this PhD research and they do not investigate how over fitting may affect the prediction of an independent sample. They therefore use a combination of the R SPLS function with unrestricted (Kappa and Eta) and the R mixOmics function which specifies a fixed number of variables to keep in the model.

This PhD research found that when modelling a single Y variable (Larsen score), a second component was not required even when the first component was restricted to very few variables. Therefore, investigating Kappa and Eta (for the number of components & number of variables) is not anticipated to help with model fitting and the bigger issue appears to be the avoidance of over fitting of the model. In this research, over fitting was addressed by the creation of a 'two stage' or 'three stage average rank' method. There was also evidence that optimising kappa and eta, when data is split into smaller blocks prior to combining into a higher level model, may try to optimise the model too soon and lead to exclusion of potentially important variables (as too few variables are carried forward to the higher level model). The proposed 'average rank' method allows many variables to be carried forward from the lower level models and optimises the model fit only on the higher level model. Whilst the research could have investigated using tuning parameters on the higher level model, the 'average rank' method appears to prevent over fitting of the model when used in combination with the 'two-stage' or 'three-stage' model fitting strategy.

Le Cao et al. (2011) have data on 525,000 SNPs however they pre-filter the SNPs removing any with a minor allele frequency less than 0.05 and then randomly sample 20,000 SNPs for use in their comparison of multivariate methods for discrimination. Le Floch et al. (2012) investigate 622,534 SNPs using various methods to reduce the dimension including using the rank of the loading to select the 'top' SNPs. They suggest univariate filtering as a mandatory step to prevent over fitting. Whilst they comment that multivariate filtering would be preferential to univariate filtering, in order to account for interactions, they do not research how this would be performed.

The research in this PhD using the different forms of validation in the 'three stage average rank' method, provides a solution to the pre-filtering problem. It is anticipated that the models created will be more transferable and robust to over fitting. As no pre-filtering is required prior to modelling, there is no loss of information. Initially, models are created under multiple internal CV to provide an order of the variables being investigated. Following on from this, variable selection is performed using separate training & test sets. The model is then tested using a third independent testing set. This results in no bias in the estimate of the prediction ability due to those predicted subjects not being used at all in model training. Whilst the method is still limited to only being applicable to the cohort of data being used, if the cohort is representative of the population, then it should be transferable to other cohorts. To date, no other method has been identified in the literature which performs variable reduction multivariately, prevents over fitting of the final models and provides an estimate of model prediction on an independent sample.

It is therefore believed that the 'three stage average rank' method implements a more suitable dimension reduction method as part of the SPLS model fitting strategy to avoid the requirement for pre-filtering which could lead to multivariate information in the data being lost. Although this research was not focused on the detection of causal SNPs, Abraham et al. (2013) used LASSO and elastic net methods to investigate 270,657 SNPs and concluded that pre-filtering limits the probability to detect the causal SNP. This supports the requirement of a multivariate variable reduction method as opposed to any univariate filtering.

Turkmen and Lin, (2012) propose a two-step method 'rPLS' which performed regularisation to reduce the SNPs followed by PLS methods. Whilst similar to the approach applied in this PhD research, the 'three stage average rank' method has the added advantage of ordering and selecting the variables for the final model using PLS methods and then being able to independently test the model fit on the set of data left out of model training. Variables excluded during this process will have been considered multivariately including their relationship with the Y response rather than using any univariate pre-filtering methods.

One limitation of the 'three stage average rank' method is that sufficient patients are required to be able to separate the sample into three groups, each with enough representative data to both create and test the model adequately. Further research would be required to investigate the optimum split of the data. A 40% 'variable ordering training set', a 40% 'variable selection training set' and a 20% independent test set was used in this research, however, simulations to investigate the model variability subject to different splits would be beneficial.

A further limitation is that the GoRA data is a cross sectional sample, measured on relatively heterogeneous patients with varied disease duration and different treatment regimens which have been administered over varying lengths of duration. This introduces variation into the data which cannot be explained by the measurements recorded. A patient susceptible to severe disease but administered with an aggressive treatment regimen may quickly have their progression stopped. Another patient with the same susceptibility to severe disease may remain undiagnosed and untreated for years and thus experience rapid progression. Unfortunately, these differences cannot be accounted for in the analysis of this data. Applying these techniques to a more homogenous sample which has details of treatments administered may be able to detect smaller SNP effects due to less unexplained variation in the data.

10.5. Areas for future research

The 'three stage average rank' method used splits of the data into a 40% 'variable ordering training set', a 40% 'variable selection training set' and a 20% independent test set. This created a model which avoiding over fitting. Within the 40% 'variable ordering training set', internal CV models were fitted which split the data into folds which were run a number of times. For each of these runs, each patient is left out of the training data once until each subject receives a predicted value. Rather than using CV, Wehrens et al. (2011) suggest leaving a random 30% of the samples out each time (similar to the 20% left out by the 5-fold CV) and also leaving a random 50% of the variables out each time and repeating this 100 times. Variables are ranked by the size of their coefficients and the final model is determined by the number of times a variable was selected in the top 10% out of the number of times it was actually included in the model. Due to time constraints it was not possible to use this approach on the GoRA data. However, if a suitable dataset could be obtained which ideally had less variation with regards to each subjects disease duration, then exploration into this methodology used alongside the SPLS methods developed in this research would be recommended.

Daetwyler et al. (2013) suggest presenting the mean and STD of the correlation to demonstrate how varied the model's prediction ability is, based on different splits of the data. This in turn can give insight as to whether the training data sample size is sufficient to form a stable model. It would be recommended to repeat the 'three stage average rank' method using different splits of the 40%, 40% and 20% sets. This would provide an estimate of the variability around the correlation between the actual and predicted values formed on different splits of the data. It would also provide information regarding whether the same variables are selected when different splits of the data are used. If different variables are selected then this indicates that the sample size may not be large enough to create a robust model and the model may not be transferable to other cohorts. Some exploration into this was provided in section 9.5 by using two different splits of the 40%, 40% and 20% sets. Although there was some variability in the correlation achieved using different splits of the data (r=0.547 to r=0.584), the top seven variables selected were the same in both models. However, as there was some variation in the variables selected after the top seven, this indicates that the sample perhaps was not large enough to reliably identify variables with smaller predictive influence on the Larsen score. However, the results are quite reassuring and the 'three stage average rank' method was shown to be more stable than the 'two stage average rank' method.

In the review of multivariate methods, using random forest methods was originally ruled out as no studies were found using them to predict continuous traits using GWAS data. However, it may be of interest to see how they perform in comparison to the PLS research presented. In addition, exploration using structural equation modelling or other penalised methods (such as LASSO and elastic net) would also be of interest for comparison with the PLS findings. It is anticipated that these methods would suffer from the same over fitting issues observed using PLS and so the use of a type of three stage procedure may be required.

10.6. Concluding review of software used

As discussed in section 3.7.5, many software packages can be used to fit PLS analyses, however, they do have different capabilities and functionality. This section provides justification of the software used throughout this research, along with the advantages and disadvantages.

Initial choices of which software to use for analyses of this project was limited by software freely available at the University of Sheffield. Although the majority of the work was conducted using the free software R using version 2.13.1 on a standard computer, section 8.2 was performed using version 2.15.1 on a Linux based high performance computing cluster 'Iceberg'. This was to enable the 100 permutations to be run simultaneously on different computing clusters. The mixOmics R package (González et al., 2011, Lê Cao et al., 2009) version 3.1 was used on a standard computer whereas the version 4.0-2 was used on the Linux clusters as a later version was available when the work was performed. The mixOmics SPLSCV function allows specification of the number of variables required to retain in the final model. This is substantially easier when attempting to reduce 325482 SNPs to just the most predictive than using the SPLS package (Chun and Keles, 2010) which requires optimisation of tuning parameters. It would be extremely hard to optimise the tuning parameters when all of the variables are not able to be fitted in the same model at the same time. The reason for this is that a model is trying to be optimised without including all of the potentially important variables in it and depending on which order you fitted the variables into groups may result in a different suggested model. In addition, as it was discovered in this work that only one component was required for modelling of the Larsen score, the tuning parameter kappa would equal one, hence only the number of variables (eta) would require optimising. The number of variables in PLS even when assessed through CV is subject to over fitting. Hence, the methods developed using the 'three stage average rank' method appear more suitable to prevent over fitting than investigating optimising tuning parameters on this data.

Table 10.3 provides an overview of the functionality available in each of the PLS software packages investigated, in addition to identifying where this software was used in this thesis. SAS version 9.2 was used for data manipulation, formatting data, merging of SNPs and verifying the PLS function is correctly implemented in R. It was not investigated for the modelling of the 'GWAS SNPs' dataset because it couldn't perform SPLS and hence had no method to reduce the variables in the model. After forming a prediction model using mixOmics, quantification was required of how much of the variance explained was due to component blocks (i.e. to the genetic variables as opposed to the environment or disease/symptom duration). No function was found in R to do this and therefore the functionality of Tanagra and MATLAB were investigated. Using the MVP Toolbox (Skov et al., 2008) in MATLAB was the better software to answer this question. Tanagra only provided estimates of each variable's contribution to each component and not the contribution of a block of variables to the full prediction model.

After being awarded additional Medical Research Council funding to purchase SIMCA and COREXpress software, the project was able to investigate whether these commercial software products had better functionality than R. SIMCA had numerous benefits such that it could analyse the 'GWAS SNPs' dataset in one PLS analysis and had excellent plot capabilities. Although, it had no built in variable selection (SPLS) method, a form of variable reduction could be done by hand after extracting the loading vector from the model and ranking them. This resulted in a very similar model to the mixOmics model (section 7.3). SIMCA was much faster than R not only because it ran

all of the 'GWAS SNPs' dataset in one model (instead of 40 blocks) but also because it only required one run of seven fold CV. R was investigated using just one run, however, the results obtained were less in agreement with the 10 runs in R and one run in SIMCA. An alternative solution to reduce model fitting time in R may be to use three runs instead of 10 runs in accordance with Long et al. (2011).

SIMCA also enabled orthogonal PLS to be explored which rotated the loading vector for a single Y variable prediction onto a simpler plane for easier visual interpretation (section 7.2). A limitation of SIMCA is that because it is not a programming language, it can take longer than R to repeat analyses on permutations of Y data, subsets of subjects or on different splits of a dataset. For example, splitting the data into various 80% training and 20% test sets or using a 40% variable ordering training set, a 40% variable selection training set and a 20% independent test set.

In conclusion, mixOmics may be more desirable if doing SPLS analysis to multiple datasets or splitting the patient cohort. Once the pre-processing is done the whole process can be automated and the code can be re-run on different datasets. However, SIMCA is preferred if modelling a single dataset as there is no need to perform any pre-processing or imputation. SIMCA runs substantially quicker than R and it has better graphics.

Table 10.3 Summary of PLS functionality by software

Software	Chapter/ Section in thesis	SPLS	For CV required to impute	Variance partitioning	Number X's able to model	Additional functions
R mixOmics v3.1 and 4.0-2[a]	4, 5, 6, 7, 8, 9	✓	✓	×	<12000	
R SPLS v 2.1-0[a]	Not used	✓	✓	×	<12000	
SAS V 9.2[b]	2.2 and 3.7.5	By hand	UNK	×	UNK	
Tanagra v1.4.44 released May 14 2012[c]	4.3.4.1	By hand	UNK	✓ (component contribution)	UNK	
MATLAB® version 7.13 R2011b[d]	4.3.4.2, 4.4.4 and 6.5.3	By hand	UNK	✓ (multi-block)	UNK	
SIMCA® Umetrics AB Version 13.0[e]	7	By hand	×	×	>325482	Orthogonal PLS, plots
CoreExpress v1.0[f]	6.4.4	✓	×	×	>24374	CCR

a: R Foundation for Statistical Computing v2.13.1 or v2.15.1 (Vienna, Austria)
b: SAS System for Windows Copyright © 2002-2008 (SAS Institute Inc. Cary NC USA)
c: (Rakotomalala, 2005)
d: The MathWorks Inc., Natick, MA, USA
e: (Eriksson et al., 2006a, Eriksson et al., 2006b)
f: (Magidson, 2011)
UNK: Unknown and not investigated during this thesis.

COExpress had SPLS functionality similar to the original version ('percentage fold' method) of mixOmics in R using multiple runs of CV. The restrictions in the software of no programming functionality and requiring selection of variables based on the numbers of folds and runs instead of using average ranked loadings, meant that the model obtained (using PLS or CCR in COExpress) was very different to that observed using the amended mixOmics version in R or SIMCA.

COREXpress was capable of modelling entire chromosomes at once although modelling the entire GWAS in one go was not attempted.

The software PLINK was used as the original 'GWAS SNPs' dataset file was stored in this common genetic data format. PLINK also has its own imputation methods which were explored on this project. Gtool (created by Colin Freeman and Jonathan Marchini and available from: http://www.well.ox.ac.uk/~cfreeman/ software/gwas/gtool.html), was used to perform the transfer of files across various data formats from PLINK to IMPUTE2 to SAS and to R. IMPUTE2 (Howie et al., 2009) was used to perform haplotype based imputation.

10.7. Summary

In conclusion, modelling strategies have been developed to be able to use SPLS on very large data problems enabling variable ordering, selection and testing. Running times using the 'average rank' method are more practical and efficient compared to the commonly used 'percentage fold' method or other linear regression based methods. The methods developed could be used on any size of dataset and have been shown to identify known predictors of RA severity. It is anticipated that if a more homogenous cohort could be found, or the sample size increased, then the methods may be suitable for identifying smaller effects. As it was not possible to account for treatments administered in this analysis, this may have reduced the ability to detect genetic predictors of RA severity. Further research is required to apply the statistical methods developed in this research to a larger cohort. This would enable a prediction model to be developed, which could be used in clinic at disease diagnosis, to estimate the patient's future risk of severe RA disease.

11. References

ABDI, H. 2010. Partial least squares regression and projection on latent structure regression (PLS Regression) [online] Wiley interdisciplinary reviews. Wiley. Available from: http://wires.wiley.com/WileyCDA/WiresArticle/wisId-WICS51.html [Accessed 18th April 2011].

ABDI, H., CHIN, W., ESPOSITO VINZI, V., RUSSOLILLO, G. & TRINCHERA, L. 2013. *New Perspectives in Partial Least Squares and Related Methods*, Springer Verlag.

ABRAHAM, G., KOWALCZYK, A., ZOBEL, J. & INOUYE, M. 2012. SparSNP: Fast and memory-efficient analysis of all SNPs for phenotype prediction. *Bmc Bioinformatics,* 13.

ABRAHAM, G., KOWALCZYK, A., ZOBEL, J. & INOUYE, M. 2013. Performance and Robustness of Penalized and Unpenalized Methods for Genetic Prediction of Complex Human Disease. *Genetic Epidemiology,* 37**,** 184-195.

AGARWAL, S. K. 2011. Core Management Principles in Rheumatoid Arthritis to Help Guide Managed Care Professionals. *Journal of Managed Care Pharmacy,* 17**,** S3-S8.

AHLMEN, M., SVENSSON, B., ALBERTSSON, K., FORSLIND, K., HAFSTROM, I. & GRP, B. S. 2010. Influence of gender on assessments of disease activity and function in early rheumatoid arthritis in relation to radiographic joint damage. *Annals of the Rheumatic Diseases,* 69**,** 230-233.

AJEGANOVA, S., ANDERSSON, M. L., HAFSTROM, I. & BARFOT STUDY, G. 2013. Association of obesity with worse disease severity in rheumatoid arthritis as well as with comorbidities: A long-term followup from disease onset. *Arthritis Care & Research,* 65**,** 78-87.

ALEXANDER, D. H. & LANGE, K. 2011. Stability Selection for Genome-Wide Association. *Genetic Epidemiology,* 35**,** 722-728.

ANDERSSON, C. A. & BRO, R. 2000. The N-way Toolbox for MATLAB. *Chemometrics and Intelligent Laboratory Systems,* 52.

ARNETT, F. C., EDWORTHY, S. M., BLOCH, D. A., MCSHANE, D. J., FRIES, J. F., COOPER, N. S., HEALEY, L. A., KAPLAN, S. R., LIANG, M. H., LUTHRA, H. S., MEDSGER, T. A., MITCHELL, D. M., NEUSTADT, D. H., PINALS, R. S., SCHALLER, J. G., SHARP, J. T., WILDER, R. L. & HUNDER, G. G. 1988. The American-Rheumatism-Association 1987 revised criteria for the classification of Rheumatoid Arthritis. *Arthritis and Rheumatism,* 31**,** 315-324.

AYERS, K. L. & CORDELL, H. J. 2010. SNP Selection in Genome-Wide and Candidate Gene Studies via Penalized Logistic Regression. *Genetic Epidemiology,* 34.

BAJPAI, U. D., SWAINSON, L. A., MOLD, J. E., GRAF, J. D., IMBODEN, J. B. & MCCUNE, J. M. 2012. A functional variant in FCRl3 is associated with higher fc receptor-like 3 expression on T cell subsets and rheumatoid arthritis disease activity. *Arthritis and Rheumatism,* 64**,** 2451-2459.

BAKER, J. F., GEORGE, M., BAKER, D. G., TOEDTER, G., VON FELDT, J. M. & LEONARD, M. B. 2011. Associations between body mass, radiographic joint damage, adipokines and risk factors for bone loss in rheumatoid arthritis. *Rheumatology,* 50**,** 2100-2107.

BARRETT, J. C., FRY, B., MALLER, J. & DALY, M. J. 2005. Haploview: analysis and visualization of LD and haplotype maps. *Bioinformatics,* 21**,** 263-265.

BARTON, A., PLATT, H., SALWAY, F., SYMMONS, D., LUNT, M., WORTHINGTON, J. & SILMAN, A. 2004. Polymorphisms in the mannose binding lectin (MBL) gene are not associated with radiographic erosions in rheumatoid or inflammatory polyarthritis. *Journal of Rheumatology,* 31**,** 442-447.

BARTON, A., THOMSON, W., KE, X., EYRE, S., HINKS, A., BOWES, J., PLANT, D., GIBBONS, L. J., WILSON, A. G., BAX, D. E., MORGAN, A. W., EMERY, P., STEER, S., HOCKING, L., REID, D. M., WORDSWORTH, P., HARRISON, P., WORTHINGTON, J., WELLCOME TRUST CASE CONTROL, C., YEAR, C. & CONSORTIUM, B. 2008. Rheumatoid arthritis susceptibility loci at chromosomes 10p15, 12q13 and 22q13. *Nature Genetics,* 40**,** 1156-1159.

BJORK, M., GERDLE, B., THYBERG, I. & PEOLSSON, M. 2008. Multivariate relationships between pain intensity and other aspects of health in rheumatoid arthritis - cross sectional and five year longitudinal analyses (the Swedish TIRA project). *Disability and Rehabilitation,* 30**,** 1429-1438.

BOHANEC GRABAR, P., LOGAR, D., TOMSIC, M., ROZMAN, B. & DOLZAN, V. 2009. Genetic polymorphisms of glutathione S-transferases and disease activity of rheumatoid arthritis. *Clin Exp Rheumatol,* 27**,** 229-36.

BOINI, S. & GUILLEMIN, F. 2001. Radiographic scoring methods as outcome measures in rheumatoid arthritis: properties and advantages. *Annals of the Rheumatic Diseases,* 60**,** 817-826.

BOULESTEIX, A.-L. 2004. PLS dimension reduction for classification with microarray data. *Stat Appl Genet Mol Biol,* 3**,** Article33.

BOULESTEIX, A.-L., BENDER, A., BERMEJO, J. L. & STROBL, C. 2012. Random forest Gini importance favours SNPs with large minor allele frequency: impact, sources and recommendations. *Briefings in Bioinformatics,* 13**,** 292-304.

BOULESTEIX, A.-L. & SAUERBREI, W. 2011. Added predictive value of high-throughput molecular data to clinical data and its validation. *Briefings in Bioinformatics,* 12**,** 215-229.

BOULESTEIX, A. L. & STRIMMER, K. 2007. Partial least squares: a versatile tool for the analysis of high-dimensional genomic data. *Briefings in Bioinformatics,* 8**,** 32-44.

BOWES, J., HO, P., FLYNN, E., ALI, F., MARZO-ORTEGA, H., COATES, L. C., WARREN, R. B., MCMANUS, R., RYAN, A. W., KANE, D., KORENDOWYCH, E., MCHUGH, N., FITZGERALD, O., PACKHAM, J., MORGAN, A. W., BRUCE, I. N. & BARTON, A. 2012. Comprehensive assessment of rheumatoid arthritis susceptibility loci in a large psoriatic arthritis cohort. *Annals of the Rheumatic Diseases,* 71**,** 1350-1354.

BRENOL, C. V., CHIES, J. A. B., BRENOL, J. C. T., MONTICIELO, O. A., FRANCISCATTO, P., BIRRIEL, F., NEVES, A. G. & XAVIER, R. M. 2009. Endothelial nitric oxide synthase T-786C polymorphism in rheumatoid arthritis: association with extraarticular manifestations. *Clinical Rheumatology,* 28**,** 201-205.

BRESNIHAN, B. 2002. Rheumatoid arthritis: principles of early treatment. *The Journal of rheumatology. Supplement,* 66**,** 9-12.

BUCHS, N., DI GIOVINE, F. S., SILVESTRI, T., VANNIER, E., DUFF, G. W. & MIOSSEC, P. 2001. IL-1B and IL-1Ra gene polymorphisms and disease severity in rheumatoid arthritis: interaction with their plasma levels. *Genes and Immunity,* 2**,** 222-228.

BUKHARI, M., LUNT, M., HARRISON, B. J., SCOTT, D. G. I., SYMMONS, D. P. M. & SILMAN, A. J. 2002. Rheumatoid factor is the major predictor of increasing severity of radiographic erosions in rheumatoid arthritis - Results from the Norfolk Arthritis Register study, a large inception cohort. *Arthritis and Rheumatism,* 46**,** 906-912.

BYLESJO, M., RANTALAINEN, M., NICHOLSON, J. K., HOLMES, E. & TRYGG, J. 2008. K-OPLS package: Kernel-based orthogonal projections to latent structures for prediction and interpretation in feature space. *Bmc Bioinformatics,* 9.

CANTAGREL, A., NAVAUX, F., LOUBET-LESCOULIE, P., NOURHASHEMI, F., ENAULT, G., ABBAL, M., CONSTANTIN, A., LAROCHE, M. & MAZIERES, B. 1999. Interleukin-1 beta, interleukin-1 receptor antagonist, interleukin-4, and interleukin-10 gene polymorphisms - Relationship to occurrence and severity of rheumatoid arthritis. *Arthritis and Rheumatism,* 42**,** 1093-1100.

CAPLAN, L., DAVIS, L. A., BRIGHT, C. M., KERR, G. S., LAZARO, D. M., KHAN, N. A., RICHARDS, J. S., JOHNSON, D. S., CANNON, G. W., REIMOLD, A. M. & MIKULS, T. R. 2013. Body mass index and the rheumatoid arthritis swollen joint count: An observational study. *Arthritis Care & Research,* 65**,** 101-106.

CASP. 2010. **Critical Appraisal Skills Programme (CASP). [online] Available from: http://www.sph.nhs.uk/what-we-do/public-health-workforce/resources/critical-appraisals-skills-programme [Accessed 15 January 2011]**.

CECCARELLI, F., PERRICONE, C., FABRIS, M., ALESSANDRI, C., IAGNOCCO, A., FABRO, C., PONTARINI, E., DE VITA, S. & VALESINI, G. 2011. Transforming growth factor beta 869C/T and interleukin 6 -174G/C polymorphisms relate to the severity and progression of bone-erosive damage detected by ultrasound in rheumatoid arthritis. *Arthritis Research & Therapy,* 13.

CHEN, J.-Y., WANG, C.-M., WU, Y.-J. J., KUO, S.-N., SHIU, C.-F., CHANG, S.-W., LIN, Y.-T., HO, H.-H. & WU, J. 2011. Disease Phenotypes and Gender Association of FCRL3 Single-Nucleotide Polymorphism-169T/C in Taiwanese Patients with Systemic Lupus Erythematosus and Rheumatoid Arthritis. *Journal of Rheumatology,* 38**,** 264-270.

CHEN, Y. F., JOBANPUTRA, P., BARTON, P., JOWETT, S., BRYAN, S., CLARK, W., FRY-SMITH, A. & BURLS, A. 2006. A systematic review of the effectiveness of adalimumab, etanercept and infliximab for the treatment of rheumatoid arthritis in adults and an economic evaluation of their cost-effectiveness. *Health Technology Assessment,* 10**,** 1-+.

CHUN, H. & KELES, S. 2010. Sparse partial least squares regression for simultaneous dimension reduction and variable selection. *Journal of the Royal Statistical Society Series B-Statistical Methodology,* 72**,** 3-25.

CONSTANTIN, A., LAUWERS-CANCES, V., NAVAUX, F., ABBAL, M., VAN MEERWIJK, J., MAZIERES, B., CAMBON-THOMSEN, A. & CANTAGREL, A. 2002. Stromelysin 1 (matrix metalloproteinase 3) and HLA-DRB1 gene polymorphisms - Association with severity and progression of rheumatoid arthritis in a prospective study. *Arthritis and Rheumatism,* 46**,** 1754-1762.

COOPER, J. D., SMYTH, D. J., SMILES, A. M., PLAGNOL, V., WALKER, N. M., ALLEN, J. E., DOWNES, K., BARRETT, J. C., HEALY, B. C., MYCHALECKYJ, J. C., WARRAM, J. H. & TODD, J. A. 2008. Meta-analysis of genome-wide association study data identifies additional type 1 diabetes risk loci. *Nature Genetics,* 40**,** 1399-1401.

CORNELIS, M. C., BAE, S. C., KIM, I. & EL-SOHEMY, A. 2010. CYP1A2 genotype and rheumatoid arthritis in Koreans. *Rheumatology International,* 30**,** 1349-1354.

COX, T. F. 2005. *An Introduction To Multivariate Data Analysis,* London, Hodder Arnould.

CROISEAU, P. & CORDELL, H. J. 2009. Analysis of North American Rheumatoid Arthritis Consortium data using a penalized logistic regression approach. *BMC proceedings,* 3 Suppl 7**,** S61-S61.

CVETKOVIC, J. T., WALLBERG-JONSSON, S., STEGMAYR, B., RANTAPAA-DAHLQVIST, S. & LEFVERT, A. K. 2002. Susceptibility for and clinical manifestations of rheumatoid arthritis are associated with polymorphisms of the TNF-alpha, IL-1 beta, and IL-lRa genes. *Journal of Rheumatology,* 29**,** 212-219.

DAETWYLER, H. D., CALUS, M. P. L., PONG-WONG, R., DE LOS CAMPOS, G. & HICKEY, J. M. 2013. Genomic Prediction in Animals and Plants: Simulation of Data, Validation, Reporting, and Benchmarking. *Genetics,* 193**,** 347-+.

DEVRIES, N., PRINSEN, C. F. M., MENSINK, E., VANRIEL, P., VANTHOF, M. A. & VANDEPUTTE, L. B. A. 1993. A T-CELL RECEPTOR BETA-CHAIN VARIABLE REGION POLYMORPHISM ASSOCIATED WITH RADIOGRAPHIC PROGRESSION IN RHEUMATOID-ARTHRITIS. *Annals of the Rheumatic Diseases,* 52**,** 327-331.

DIMAURO, C., STERI, R., PINTUS, M. A., GASPA, G. & MACCIOTTA, N. P. P. 2011. Use of partial least squares regression to predict single nucleotide polymorphism marker genotypes when some animals are genotyped with a low-density panel. *Animal,* 5**,** 833-837.

DORR, S., LECHTENBOHMER, N., RAU, R., HERBORN, G., WAGNER, U., MULLER-MYHSOK, B., HANSMANN, I. & KEYSZER, G. 2004. Association of a specific haplotype across the genes MMP1 and MMP3 with radiographic joint destruction in rheumatoid arthritis. *Arthritis Research & Therapy,* 6**,** R199-R207.

EASTMENT, H. T. & KRZANOWSKI, W. J. 1982. CROSS-VALIDATORY CHOICE OF THE NUMBER OF COMPONENTS FROM A PRINCIPAL COMPONENT ANALYSIS. *Technometrics,* 24**,** 73-77.

EFRON, B. & GONG, G. 1983. A LEISURELY LOOK AT THE BOOTSTRAP, THE JACKKNIFE, AND CROSS-VALIDATION. *American Statistician,* 37**,** 36-48.

ENEVOLD, C., RADSTAKE, T. R. D., COENEN, M. J. H., FRANSEN, J., TOONEN, E. J. M., BENDTZEN, K. & VAN RIEL, P. 2010. Multiplex Screening of 22 Single-Nucleotide Polymorphisms in 7 Toll-like Receptors: An Association Study in Rheumatoid Arthritis. *Journal of Rheumatology,* 37**,** 905-910.

ERIKSSON, L., ANTTI, H., GOTTFRIES, J., HOLMES, E., JOHANSSON, E., LINDGREN, F., LONG, I., LUNDSTEDT, T., TRYGG, J. & WOLD, S. 2004. Using chemometrics for navigating in the large data sets of genomics, proteomics, and metabonomics (gpm). *Analytical and Bioanalytical Chemistry,* 380**,** 419-429.

ERIKSSON, L., JOHANSSON, E., KETTANEH-WOLD, N., TRYGG, J., WIKSTRÖM, C. & WOLD, S. 2006a. *Multi-and Megavariate Data Analysis, Part I, Basic Principals and Applications*, Umetrics Academy.

ERIKSSON, L., JOHANSSON, E., KETTANEH-WOLD, N., TRYGG, J., WIKSTRÖM, C. & WOLD, S. 2006b. *Multi-and Megavariate Data Analysis, Part II, Advanced Applications and Method Extensions*, Umetrics Academy.

EYRE, S., BOWES, J., DIOGO, D., LEE, A., BARTON, A., MARTIN, P., ZHERNAKOVA, A., STAHL, E., VIATTE, S., MCALLISTER, K., AMOS, C. I., PADYUKOV, L., TOES, R. E. M., HUIZINGA, T. W. J., WIJMENGA, C., TRYNKA, G., FRANKE, L., WESTRA, H.-J., ALFREDSSON, L., HU, X., SANDOR, C., DE BAKKER, P. I. W., DAVILA, S., KHOR, C. C., HENG, K. K., ANDREWS, R., EDKINS, S., HUNT, S. E., LANGFORD, C., SYMMONS, D., CONCANNON, P., ONENGUT-GUMUSCU, S., RICH, S. S., DELOUKAS, P., GONZALEZ-GAY, M. A., RODRIGUEZ-RODRIGUEZ, L., ARLSETIG, L., MARTIN, J., RANTAPAA-DAHLQVIST, S., PLENGE, R. M., RAYCHAUDHURI, S., KLARESKOG, L., GREGERSEN, P. K., WORTHINGTON, J., BIOL RHEUMATOID ARTHRIT GENETICS, G. & WELLCOME TRUST CASE CONTROL, C. 2012. High-density genetic mapping identifies new susceptibility loci for rheumatoid arthritis. *Nature Genetics,* 44**,** 1336-1340.

EYRE, S., FLYNN, E., MARTIN, P., HINKS, A., WILSON, A. G., MORGAN, A. W., EMERY, P., STEER, S., HOCKING, L. J., REID, D. M., HARRISON, P., WORDSWORTH, P., THOMSON, W., WORTHINGTON, J. & BARTON, A. 2010a. No evidence for association of the KLF12 gene with rheumatoid arthritis in a large UK cohort. *Annals of the Rheumatic Diseases,* 69**,** 1407-U189.

EYRE, S., HINKS, A., BOWES, J., FLYNN, E., MARTIN, P., WILSON, A. G., MORGAN, A. W., EMERY, P., STEER, S., HOCKING, L. J., REID, D. M., HARRISON, P., WORDSWORTH, P., THOMSON, W., WORTHINGTON, J., BARTON, A. & CONSORTIUM, B. 2010b. Overlapping genetic susceptibility variants between three autoimmune disorders: rheumatoid arthritis, type 1 diabetes and coeliac disease. *Arthritis Research & Therapy,* 12.

FAROUK, H. M., MANSOUR, H. E., RAHMAN, S. A., MOSTAFA, A. A., SHAMY, H. A. & ZAROUK, W. A. 2009. Effect of the human leukocyte antigen HLA-DRB1 and anti-cyclic citrullinated peptide on the outcome of rheumatoid arthritis patients. *Brazilian Journal of Medical and Biological Research,* 42**,** 831-838.

FENG, Z. Z., YANG, X., SUBEDI, S. & MCNICHOLAS, P. D. 2012. The LASSO and Sparse Least Squares Regression Methods for SNP Selection in Predicting Quantitative Traits. *Ieee-Acm Transactions on Computational Biology and Bioinformatics,* 9**,** 629-636.

FIRESTEIN, G. S., PANAYI, G. S. & WOLLHEIM, F. A. 2006. *Rheumatoid arthritis,* Oxford, Oxford University Press.

FONVILLE, J. M., RICHARDS, S. E., BARTON, R. H., BOULANGE, C. L., EBBELS, T. M. D., NICHOLSON, J. K., HOLMES, E. & DUMAS, M. E. 2010. The evolution of partial least squares models and related chemometric approaches in metabonomics and metabolic phenotyping. *Journal of Chemometrics,* 24**,** 636-649.

FORNELL, C. & BOOKSTEIN, F. L. 1982. 2 STRUCTURAL EQUATION MODELS - LISREL AND PLS APPLIED TO CONSUMER EXIT-VOICE THEORY. *Journal of Marketing Research,* 19**,** 440-452.

FRANK, I. E. & FRIEDMAN, J. H. 1993. A STATISTICAL VIEW OF SOME CHEMOMETRICS REGRESSION TOOLS. *Technometrics,* 35**,** 109-135.

GAMBHIR, D., LAWRENCE, A., AGGARWAL, A., MISRA, R., MANDAL, S. K. & NAIK, S. 2010. Association of tumor necrosis factor alpha and IL-10 promoter polymorphisms with rheumatoid arthritis in North Indian population. *Rheumatology International,* 30**,** 1211-1217.

GARTHWAITE, P. H. 1994. AN INTERPRETATION OF PARTIAL LEAST-SQUARES. *Journal of the American Statistical Association,* 89**,** 122-127.

GELADI, P. & KOWALSKI, B. R. 1986. PARTIAL LEAST-SQUARES REGRESSION - A TUTORIAL. *Analytica Chimica Acta,* 185**,** 1-17.

GENEVAY, S., DI GIOVINE, F. S., PERNEGER, T. V., SILVESTRI, T., STINGELIN, S., DUFF, G. & GUERNE, P. A. 2002. Association of interleukin-4 and interleukin-1B gene variants with Larsen score progression in rheumatoid arthritis. *Arthritis & Rheumatism-Arthritis Care & Research,* 47**,** 303-309.

GENG, Y., ZHOU, W. & ZHANG, Z.-L. 2012. A comparative study on the diversity of clinical features between the sero-negative and sero-positive rheumatoid arthritis patients. *Rheumatology International,* 32**,** 3897-3901.

GENIN, E., BABRON, M. C., MCDERMOTT, M. F., MULCAHY, B., WALDRON-LYNCH, F., ADAMS, C., CLEGG, D. O., WARD, R. H., SHANAHAN, F., MOLLOY, M. G., O'GARA, F. & CLERGET-DARPOUX, F. 1998. Modelling the major histocompatibility complex susceptibility to RA using the MASC method. *Genetic Epidemiology,* 15**,** 419-430.

GENNEBACK, N., MALM, L., HELLMAN, U., WALDENSTROM, A. & MORNER, S. 2013. Using OPLS-DA to find new hypotheses in vast amounts of gene expression data - Studying the progression of cardiac hypertrophy in the heart of aorta ligated rat. *Gene,* 522**,** 27-36.

GONZÁLEZ, I., LÊ CAO, K.-A. & DÉJEAN, S. 2011. mixOmics: Omics Data Integration Project. URL: http://www.math.univ-toulouse.fr/~biostat/mixOmics/.

GOSSEC, L., BARO-RIBA, J., BOZONNAT, M. C., DAURES, J. P., SANY, J., ELIAOU, J. F. & COMBE, B. 2005. Influence of sex on disease severity in patients with rheumatoid arthritis. *Journal of Rheumatology,* 32**,** 1448-1451.

GOTTENBERG, J.-E., DAYER, J.-M., LUKAS, C., DUCOT, B., CHIOCCHIA, G., CANTAGREL, A., SARAUX, A., ROUX-LOMBARD, P. & MARIETTE, X. 2012. Serum IL-6 and IL-21 are associated with markers of B cell activation and structural progression in early rheumatoid arthritis: results from the ESPOIR cohort. *Annals of the Rheumatic Diseases,* 71**,** 1243-1248.

GOURRAUD, P. A., BOYER, J. F., BARNETCHE, T., ABBAL, M., CAMBON-THOMSEN, A., CANTAGREL, A. & CONSTANTIN, A. 2006. New classification of HLA-DRB1 alleles differentiates predisposing and protective alleles for rheumatoid arthritis structural severity. *Arthritis and Rheumatism,* 54**,** 593-599.

GRABAR, P. B., ROJKO, S., LOGAR, D. & DOLZAN, V. 2010. Genetic determinants of methotrexate treatment in rheumatoid arthritis patients: a study of polymorphisms in the adenosine pathway. *Annals of the Rheumatic Diseases,* 69**,** 931-U186.

GRAESSLER, J., VERLOHREN, M., GRAESSLER, A., ZEISSIG, A., KUHLISCH, E., KOPPRASCH, S. & SCHROEDER, H. E. 2005. Association of chondromodulin-II Val58Ile polymorphism with radiographic joint destruction in rheumatoid arthritis. *Journal of Rheumatology,* 32**,** 1654-1661.

GRAUDAL, N. 2004. The natural history and prognosis of rheumatoid arthritis: association of radiographic outcome with process variables, joint motion and immune proteins. *Scand J Rheumatol Suppl,* 118**,** 1-38.

GRAUDAL, N. A., MADSEN, H. O., TARP, U., SVEJGAARD, A., JURIK, A. G., GRAUDAL, H. K. & GARRED, P. 2000. The association of variant mannose-binding lectin genotypes with radiographic outcome in rheumatoid arthritis. *Arthritis and Rheumatism,* 43**,** 515-521.

GREGERSEN, P. K., SILVER, J. & WINCHESTER, R. J. 1987. THE SHARED EPITOPE HYPOTHESIS - AN APPROACH TO UNDERSTANDING THE MOLECULAR-GENETICS OF SUSCEPTIBILITY TO RHEUMATOID-ARTHRITIS. *Arthritis and Rheumatism,* 30**,** 1205-1213.

GUPTA, B., AGRAWAL, C., RAGHAV, S. K., DAS, S. K., DAS, R. H., CHATURVEDI, V. P. & DAS, H. R. 2005. Association of mannose-binding lectin gene (MBL2) polymorphisms with rheumatoid arthritis in an Indian cohort of case-control samples. *Journal of Human Genetics,* 50**,** 583-591.

GYETVAI, A., SZEKANECZ, Z., SOOS, L., SZABO, Z., FEKETE, A., KAPITANY, A., TEODORESCU, M., SIPKA, S., SZEGEDI, G. & LAKOS, G. 2010. New classification of the shared epitope in rheumatoid arthritis: impact on the production of various anti-citrullinated protein antibodies. *Rheumatology,* 49**,** 25-33.

HALVORSEN, E. H., HAAVARDSHOLM, E. A., POLLMANN, S., BOONEN, A., VAN DER HEIJDE, D., KVIEN, T. K. & MOLBERG, O. 2009. Serum IgG antibodies to peptidylarginine deiminase 4 predict radiographic progression in patients with rheumatoid arthritis treated with tumour necrosis factor-alpha blocking agents. *Annals of the Rheumatic Diseases,* 68**,** 249-252.

HAMMER, H. B., ODEGARD, S., FAGERHOL, M. K., LANDEWE, R., VAN DER HEIJDE, D., UHLIG, T., MOWINCKEL, P. & KVIEN, T. K. 2007. Calprotectin (a major leucocyte protein) is strongly and independently correlated with joint inflammation and damage in rheumatoid arthritis. *Annals of the Rheumatic Diseases,* 66**,** 1093-1097.

HAN, S. W., SA, K. H., KIM, S. I., LEE, S. I., PARK, Y. W., LEE, S. S., YOO, W. H., KANG, J. Y., SOE, J. S., NAM, E. J., LEE, J., PARK, J. Y. & KANG, Y. M. 2012a. FCRL3 gene polymorphisms contribute to the radiographic severity rather than susceptibility of rheumatoid arthritis. *Human Immunology,* 73**,** 537-542.

HAN, S. W., SA, K. H., KIM, S. I., LEE, S. I., PARK, Y. W., LEE, S. S., YOO, W. H., SOE, J. S., NAM, E. J., LEE, J., PARK, J. Y. & KANG, Y. M. 2012b. CCR5 gene polymorphism is a genetic risk factor for radiographic severity of rheumatoid arthritis. *Tissue Antigens,* 80**,** 416-423.

HARRELL, F. E. 2001. *Regression Modeling Strategies: With applications to linear models, logistic regression, and survival analysis,* New York, USA, Springer.

HARRIS, M. L., DARRAH, E., LAM, G. K., BARTLETT, S. J., GILES, J. T., GRANT, A. V., GAO, P., SCOTT, W. W., EL-GABALAWY, H., CASCIOLA-ROSEN, L., BARNES, K. C., BATHON, J. M. & ROSEN, A. 2008. Association of autoimmunity to peptidyl arginine deiminase type 4 with genotype and disease severity in rheumatoid arthritis. *Arthritis and Rheumatism,* 58**,** 1958-1967.

HARRISON, P., POINTON, J. J., CHAPMAN, K., RODDAM, A. & WORDSWORTH, B. P. 2008. Interleukin-1 promoter region polymorphism role in rheumatoid arthritis: a meta-analysis of IL-1B-511A/G variant reveals association with rheumatoid arthritis. *Rheumatology,* 47**,** 1768-1770.

HASTIE, T., TIBSHIRANI, R. & FRIEDMAN, J. 2009. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, Springer.

HINTZE, J. 2008. PASS 2008. NCSS, LLC. Kaysville, Utah. www.ncss.com.

HOCHBERG, M. C. 1981. ADULT AND JUVENILE RHEUMATOID-ARTHRITIS - CURRENT EPIDEMIOLOGIC CONCEPTS. *Epidemiologic Reviews,* 3**,** 27-44.

HOPPE, B., HAUPL, T., EGERER, K., GRUBER, R., KIESEWETTER, H., SALAMA, A., BURMESTER, G. R. & DORNER, T. 2009. Influence of peptidylarginine deiminase type 4 genotype and shared epitope on clinical characteristics and autoantibody profile of rheumatoid arthritis. *Annals of the Rheumatic Diseases,* 68**,** 898-903.

HOWIE, B. N., DONNELLY, P. & MARCHINI, J. 2009. A Flexible and Accurate Genotype Imputation Method for the Next Generation of Genome-Wide Association Studies. *Plos Genetics,* 5.

HUANG, C. M., TSAI, F. J., WU, J. Y. & WU, M. C. 2001. Interleukin-1 beta and interleukin-1 receptor antagonist gene polymorphisms in rheumatoid arthritis. *Scandinavian Journal of Rheumatology,* 30**,** 225-228.

HUIZINGA, T. W. J., AMOS, C. I., VAN DER HELM-VAN MIL, A. H. M., CHEN, W., VAN GAALEN, F. A., JAWAHEER, D., SCHREUDER, G. M. T., WENER, M., BREEDVELD, F. C., AHMAD, N., LUM, R. F., DE VRIES, R. R. P., GREGERSEN, P. K., TOES, R. E. M. & CRISWELL, L. A. 2005. Refining the

complex rheumatoid arthritis phenotype based on specificity of the HLA-DRB1 shared epitope for antibodies to citrullinated proteins. *Arthritis and Rheumatism,* 52**,** 3433-3438.
HUIZINGA, T. W. J., KEIJSERS, V., YANNI, G., HALL, M., RAMAGE, W., LANCHBURY, J., PITZALIS, C., DROSSAERS-BAKKER, W. K., WESTENDORP, R. G. J., BREEDVELD, F. C., PANAYI, G. & VERWEIJ, C. L. 2000. Are differences in interleukin 10 production associated with joint damage? *Rheumatology,* 39**,** 1180-1188.
IBN YACOUB, Y., AMINE, B., LAATIRIS, A. & HAJJAJ-HASSOUNI, N. 2012a. Rheumatoid factor and antibodies against citrullinated peptides in Moroccan patients with rheumatoid arthritis: association with disease parameters and quality of life. *Clinical Rheumatology,* 31**,** 329-334.
IBN YACOUB, Y., AMINE, B., LAATIRIS, A., WAFKI, F., ZNAT, F. & HAJJAJ-HASSOUNI, N. 2012b. Prevalence of overweight in Moroccan patients with rheumatoid arthritis and its relationships with disease features. *Clinical Rheumatology,* 31**,** 479-482.
INNALA, L., KOKKONEN, H., ERIKSSON, C., JIDELL, E., BERGLIN, E. & RANTAPAA-DAHLQVIST, S. 2008. Antibodies against mutated citrullinated vimentin are a better predictor of disease activity at 24 months in early rheumatoid arthritis than antibodies against cyclic citrullinated peptides. *Journal of Rheumatology,* 35**,** 1002-1008.
IP, W. K., LAU, Y. L., CHAN, S. Y., MOK, C. C., CHAN, D., TONG, K. K. & LAU, C. S. 2000. Mannose-binding lectin and rheumatoid arthritis in southern Chinese. *Arthritis and Rheumatism,* 43**,** 1679-1687.
IRIGOYEN, P., LEE, A. T., WENER, M. H., LI, W. T., KERN, M., BATLIWALLA, F., LUM, R. F., MASSAROTTI, E., WEISMAN, M., BOMBARDIER, C., REMMERS, E. F., KASTNER, D. L., SELDIN, M. F., CRISWELL, L. A. & GREGERSEN, P. K. 2005. Regulation of anti-cyclic citrullinated peptide antibodies in rheumatoid arthritis - Contrasting effects of HLA-DR3 and the shared epitope alleles. *Arthritis and Rheumatism,* 52**,** 3813-3818.
ISAACS, J. D. & MORELAND, L. W. 2002. *Rheumatoid arthritis*, Oxford : Health Press.
JACOBSEN, S., GARRED, P., MADSEN, H. O., HEEGAARD, N. H. H., HETLAND, M. L., STENGAARD-PEDERSEN, K., JUNKER, P., LOTTENBURGER, T., ELLINGSEN, T., ANDERSEN, L. S., HANSEN, I., SKJODT, H., PEDERSEN, J. K., LAURIDSEN, U. B., SVENDSEN, A. J., TARP, U., PODENPHANT, J., LINDEGAARD, H., VESTERGAARD, A., OSTERGAARD, M. & HORSLEV-PETERSEN, K. 2009. Mannose-Binding Lectin Gene Polymorphisms Are Associated with Disease Activity and Physical Disability in Untreated, Anti-Cyclic Citrullinated Peptide-Positive Patients with Early Rheumatoid Arthritis. *Journal of Rheumatology,* 36**,** 731-735.
JACOBSEN, S., MADSEN, H. O., KLARLUND, M., JENSEN, T., SKJODT, H., JENSEN, K. E., SVEJGAARD, A., GARRED, P. & GRP, T. 2001. The influence of mannose binding lectin polymorphisms on disease outcome in early polyarthritis. *Journal of Rheumatology,* 28**,** 935-942.
JANSEN, A., VAN DER HORST-BRUINSMA, I. E., VAN SCHAARDENBURG, D., VAN DE STADT, R. J., DE KONING, M. & DIJKMANS, B. A. C. 2002. Rheumatoid factor and antibodies to cyclic citrullinated peptide differentiate rheumatoid arthritis from undifferentiated polyarthritis in patients with early arthritis. *Journal of Rheumatology,* 29**,** 2074-2076.
JOERG, K., WRITER, S., JUERGEN, S. & GERT, H. 2004. A higher body mass index (BMI) prevents severe radiological progression in rheumatoid artrhitis - Is there an obesity paradox in erosive arthritis. *International Journal of Obesity,* 28**,** S119.
JOHNSEN, A. K., PLENGE, R. M., BUTTY, V., CAMPBELL, C., DIEGUEZ-GONZALEZ, R., GOMEZ-REINO, J. J., SHADICK, N., WEINBLATT, M., GONZALEZ, A., GREGERSEN, P. K., BENOIST, C. & MATHIS, D. 2008. Broad analysis of IL1 polymorphism and rheumatoid arthritis. *Arthritis and Rheumatism,* 58**,** 1947-1957.
JORGENSEN, C., PICOT, M. C., BOLOGNA, C. & SANY, J. 1996. Oral contraception, parity, breast feeding, and severity of rheumatoid arthritis. *Annals of the Rheumatic Diseases,* 55**,** 94-98.
JOUVENNE, P., CHAUDHARY, A., BUCHS, N., DI GIOVINE, F. S., DUFF, G. W. & MIOSSEC, P. 1999. Possible genetic association between interleukin-1 alpha gene polymorphism and the severity of chronic polyarthritis. *European Cytokine Network,* 10**,** 33-36.

JOVEN, B., GONZALEZ, N., AGUILAR, F., SANTIAGO, B., GALINDO, M., ALCAMI, J. & PABLOS, J. L. 2005. Association between stromal cell-derived factor 1 chemokine gene variant and radiographic progression of rheumatoid arthritis. *Arthritis and Rheumatism,* 52**,** 354-356.

JUNTA, C. M., SANDRIN-GARCIA, P., FACHIN-SALTORATTO, A. L., MELLO, S. S., OLIVEIRA, R. D. R., RASSI, D. M., GIULIATTI, S., SAKAMOTO-HOJO, E. T., LOUZADA, P., DONADI, E. A. & PASSOS, G. A. S. 2009. Differential gene expression of peripheral blood mononuclear cells from rheumatoid arthritis patients may discriminate immunogenetic, pathogenic and treatment features. *Immunology,* 127**,** 365-372.

KALLBERG, H., PADYUKOV, L., PLENGE, R. M., RONNELID, J., GREGERSEN, P. K., VAN DER HELM-VAN MIL, A. H. M., TOES, R. E. M., HUIZINGA, T. W., KLARESKOG, L., ALFREDSSON, L. & EPIDEMIOLOGICAL, I. 2007. Gene-gene and gene-environment interactions involving HLA-DRB1, PTPN22, and smoking in two subsets of rheumatoid arthritis. *American Journal of Human Genetics,* 80**,** 867-875.

KARLSON, E. W., CHIBNIK, L. B., CUI, J., PLENGE, R. M., GLASS, R. J., MAHER, N. E., PARKER, A., ROUBENOFF, R., IZMAILOVA, E., COBLYN, J. S., WEINBLATT, M. E. & SHADICK, N. A. 2008. Associations between Human leukocyte antigen, PTPN22, CTLA4 genotypes and rheumatoid arthritis phenotypes of autoantibody status, age at diagnosis and erosions in a large cohort study. *Annals of the Rheumatic Diseases,* 67**,** 358-363.

KASTBOM, A., JOHANSSON, M., VERMA, D., SODERKVIST, P. & RANTAPAA-DAHLQVIST, S. 2010. CARD8 p.C10X polymorphism is associated with inflammatory activity in early rheumatoid arthritis. *Annals of the Rheumatic Diseases,* 69**,** 723-726.

KASTBOM, A., VERMA, D., ERIKSSON, P., SKOGH, T., WINGREN, G. & SODERKVIST, P. 2008. Genetic variation in proteins of the cryopyrin inflammasome influences susceptibility and severity of rheumatoid arthritis (The Swedish TIRA project). *Rheumatology,* 47**,** 415-417.

KEYSTONE, E. C., KAVANAUGH, A. F., SHARP, J. T., TANNENBAUM, H., HUA, Y., TEOH, L. S., FISCHKOFF, S. A. & CHARTASH, E. K. 2004. Radiographic, clinical, and functional outcomes of treatment with adalimumab (a human anti-tumor necrosis factor monoclonal antibody) in patients with active rheumatoid arthritis receiving concomitant methotrexate therapy - A randomized, placebo-controlled, 52-week trial. *Arthritis and Rheumatism,* 50**,** 1400-1411.

KIM, S. Y., HAN, S. W., KIM, G. W., LEE, J. M. & KANG, Y. M. 2004. TGF-beta 1 polymorphism determines the progression of joint damage in rheumatoid arthritis. *Scandinavian Journal of Rheumatology,* 33**,** 389-394.

KNEVEL, R., GRÖNDAL, G., HUIZINGA, T. W. J., VISSER, A. W., JÓNSSON, H., VÍKINGSSON, A., GEIRSSON, Á. J., STEINSSON, K. & VAN DER HELM-VAN MIL, A. H. M. 2012a. Genetic predisposition of the severity of joint destruction in rheumatoid arthritis: a population-based study. *Annals of the Rheumatic Diseases*.

KNEVEL, R., KRABBEN, A., BROUWER, E., POSTHUMUS, M. D., WILSON, A. G., LINDQVIST, E., SAXNE, T., DE ROOY, D., DAHA, N., VAN DER LINDEN, M. P. M., STOEKEN, G., VAN TOORN, L., KOELEMAN, B., TSONAKA, R., ZHERNAKOZA, A., HOUWING-DUISTERMAAT, J. J., TOES, R., HUIZINGA, T. W. J. & VAN DER HELM-VAN MIL, A. 2012b. Genetic variants in IL15 associate with progression of joint destruction in rheumatoid arthritis: a multicohort study. *Annals of the Rheumatic Diseases,* 71**,** 1651-1657.

KOCA, S. S., ETEM, E. O., ISIK, B., YUCE, H., OZGEN, M., DAG, M. S. & ISIK, A. 2010. Prevalence and significance of MEFV gene mutations in a cohort of patients with rheumatoid arthritis. *Joint Bone Spine,* 77**,** 32-35.

KURREEMAN, F. A. S., PADYUKOV, L., MARQUES, R. B., SCHRODI, S. J., SEDDIGHZADEH, M., STOEKEN-RIJSBERGEN, G., VAN DER HELM-VAN MIL, A. H. M., ALLAART, C. F., VERDUYN, W., HOUWING-DUISTERMAAT, J., ALFREDSSON, L., BEGOVICH, A. B., KLARESKOG, L., HUIZINGA, T. W. J. & TOES, R. E. M. 2007. A candidate gene approach identifies the TRAF1/C5 region as a risk factor for rheumatoid arthritis. *Plos Medicine,* 4**,** 1515-1524.

KURREEMAN, F. A. S., STAHL, E. A., OKADA, Y., LIAO, K., DIOGO, D., RAYCHAUDHURI, S., FREUDENBERG, J., KOCHI, Y., PATSOPOULOS, N. A., GUPTA, N., SANDOR, C., BANG, S.-Y., LEE, H.-S., PADYUKOV, L., SUZUKI, A., SIMINOVITCH, K., WORTHINGTON, J., GREGERSEN, P. K., HUGHES, L. B., REYNOLDS, R. J., BRIDGES, S. L., JR., BAE, S.-C., YAMAMOTO, K., PLENGE, R. M. & INVESTIGATORS, C. 2012. Use of a Multiethnic Approach to Identify Rheumatoid-Arthritis-Susceptibility Loci, 1p36 and 17q12. *American Journal of Human Genetics,* 90**,** 524-532.

KWOH, C. K., ANDERSON, L. G., GREENE, J. M., JOHNSON, D. A., O'DELL, J. R., ROBBINS, M. L., ROBERTS, W. N. J., SIMMS, R. W. & YOOD, R. A. 2002. Guidelines for the management of rheumatoid arthritis: 2002 Update. *Arthritis Rheum,* 46**,** 328-46.

KYBURZ, D., GABAY, C., MICHEL, B. A., FINCKH, A. & SCQM, R. A. 2011. The long-term impact of early treatment of rheumatoid arthritis on radiographic progression: a population-based cohort study. *Rheumatology,* 50**,** 1106-1110.

LAMAS, J. R., RODRIGUEZ-RODRIGUEZ, L., VARADE, J., LOPEZ-ROMERO, P., TORNERO-ESTEBAN, P., ABASOLO, L., URCELAY, E. & FERNANDEZ-GUTIERREZ, B. 2010. Influence of IL6R rs8192284 Polymorphism Status in Disease Activity in Rheumatoid Arthritis. *Journal of Rheumatology,* 37**,** 1579-1581.

LARD, L. R., VAN GAALEN, F. A., SCHONKEREN, J. J. M., PIETERMAN, E. J., STOEKEN, G., VOS, K., NELISSEN, R., WESTENDORP, R. G. J., HOEBEN, R. C., BREEDVELD, F. C., TOES, R. E. M. & HUIZINGA, T. W. J. 2003. Association of the-2849 interleukin-10 promoter polymorphism with autoantibody production and joint destruction in rheumatoid arthritis. *Arthritis and Rheumatism,* 48**,** 1841-1848.

LARSEN, A. 1995. HOW TO APPLY LARSEN SCORE IN EVALUATING RADIOGRAPHS OF RHEUMATOID-ARTHRITIS IN LONG-TERM STUDIES. *Journal of Rheumatology,* 22**,** 1974-1975.

LE CAO, K.-A., BOITARD, S. & BESSE, P. 2011. Sparse PLS discriminant analysis: biologically relevant feature selection and graphical displays for multiclass problems. *BMC bioinformatics,* 12**,** 253-253.

LE CAO, K.-A., MARTIN, P. G. P., ROBERT-GRANIE, C. & BESSE, P. 2009. Sparse canonical methods for biological data integration: application to a cross-platform study. *Bmc Bioinformatics,* 10.

LE CAO, K.-A., ROSSOUW, D., ROBERT-GRANIE, C. & BESSE, P. 2008. A Sparse PLS for Variable Selection when Integrating Omics Data. *Statistical Applications in Genetics and Molecular Biology,* 7.

LE FLOCH, E., GUILLEMOT, V., FROUIN, V., PINEL, P., LALANNE, C., TRINCHERA, L., TENENHAUS, A., MORENO, A., ZILBOVICIUS, M., BOURGERON, T., DEHAENE, S., THIRION, B., POLINE, J.-B. & DUCHESNAY, E. 2012. Significant correlation between a set of genetic polymorphisms and a functional brain network revealed by feature selection and sparse Partial Least Squares. *Neuroimage,* 63**,** 11-24.

LEE, K. H., KIM, H. S., EL-SOHEMY, A., CORNELIS, M. C., UHM, W. S. & BAE, S. C. 2006. Cyclooxygenase-2 genotype and rheumatoid arthritis. *Journal of Rheumatology,* 33**,** 1231-1234.

LI, Z. & SILLANPAA, M. J. 2012. Overview of LASSO-related penalized regression methods for quantitative trait mapping and genomic selection. *Theoretical and Applied Genetics,* 125**,** 419-435.

LIE, B. A., VIKEN, M. K., ODEGARD, S., VAN DER HEIJDE, D., LANDEWE, R., UHLIG, T. & KVIEN, T. K. 2007. Associations between the PTPN22 1858C -> T polymorphism and radiographic joint destruction in patients with rheumatoid arthritis: results from a 10-year longitudinal study. *Annals of the Rheumatic Diseases,* 66**,** 1604-1609.

LODDER, M. C., DE JONG, Z., KOSTENSE, P. J., MOLENAAR, E. T. H., STAAL, K., VOSKUYL, A. E., HAZES, J. M. W., DIJKMANS, B. A. C. & LEMS, W. F. 2004. Bone mineral density in patients with rheumatoid arthritis: relation between disease severity and low bone mineral density. *Annals of the Rheumatic Diseases,* 63**,** 1576-1580.

LONG, N., GIANOLA, D., ROSA, G. J. M. & WEIGEL, K. A. 2011. Dimension reduction and variable selection for genomic selection: application to predicting milk yield in Holsteins. *Journal of Animal Breeding and Genetics,* 128**,** 247-257.

LUBBE, S., TIKLY, M., VAN DER MERWE, L., HODKINSON, B. & RAMSAY, M. 2008. Interleukin-1 receptor antagonist gene polymorphisms are associated with disease severity in Black South Africans with rheumatoid arthritis. *Joint Bone Spine,* 75**,** 422-425.

LUNDSTROM, E., HARTSHORNE, T., LI, K., LINDBLAD, S., WICK, M. C., BENGTSSON, C., ALFREDSSON, L., KLARESKOG, L. & PADYUKOV, L. 2011. Effects of GSTM1 in Rheumatoid Arthritis; Results from the Swedish EIRA study. *Plos One,* 6.

LÊ CAO, K.-A., I, G. & S, D. 2009. integrOmics: an R package to unravel relationships between two omics data sets. *Bioinformatics,* 25 (21)**,** 2855-2856. ***NOTE: the package 'integrOmics' has been renamed 'mixOmics'.***

MACCHIONI, P., NICOLI, D., CASALI, B., CATANOSO, M., FARNETTI, E., BOIARDI, L. & SALVARANI, C. 2007. The codon 72 polymorphic variants of p53 in Italian rheumatoid arthritis patients. *Clinical and Experimental Rheumatology,* 25**,** 416-421.

MACGREGOR, A. J., SNIEDER, H., RIGBY, A. S., KOSKENVUO, M., KAPRIO, J., AHO, K. & SILMAN, A. J. 2000. Characterizing the quantitative genetic contribution to rheumatoid arthritis using data from twins. *Arthritis and Rheumatism,* 43**,** 30-37.

MACKIE, S. L., TAYLOR, J. C., MARTIN, S. G., CHOKKALINCONSORTIUM, U. K. R. A. G., WORDSWORTH, P., STEER, S., WILSON, A. G., WORTHINGTON, J., EMERY, P., BARRETT, J. H., MORGAN, A. W. & CONSORTIUM, Y. 2012a. A spectrum of susceptibility to rheumatoid arthritis within HLA-DRB1: stratification by autoantibody status in a large UK population. *Genes and Immunity,* 13**,** 120-128.

MACKIE, S. L., TAYLOR, J. C., TWIGG, S., MARTIN, S. G., STEER, S., WORTHINGTON, J., BARTON, A., WILSON, A. G., HOCKING, L., YOUNG, A., EMERY, P., BARRETT, J. H. & MORGAN, A. W. 2012b. Relationship between area-level socio-economic deprivation and autoantibody status in patients with rheumatoid arthritis: multicentre cross-sectional study. *Annals of the Rheumatic Diseases,* 71**,** 1640-1645.

MAEHLEN, M. T., NORDANG, G. B., SYVERSEN, S. W., VAN DER HEIJDE, D. M., KVIEN, T. K., UHLIG, T. & LIE, B. A. 2011. FCRL3-169C/C Genotype Is Associated with Anti-citrullinated Protein Antibody-positive Rheumatoid Arthritis and with Radiographic Progression. *Journal of Rheumatology,* 38**,** 2329-2335.

MAGIDSON, J. 2011. COREXpress Users Guide: Manual for COREXpress. Belmont, MA: Statistical Innovations Inc.

MARINOU, I., HEALY, J., MEWAR, D., MOORE, D. J., DICKSON, M. C., BINKS, M. H., MONTGOMERY, D. S., WALTERS, K. & WILSON, A. G. 2007. Association of interleukin-6 and interleukin-10 genotypes with radiographic damage in rheumatoid arthritis is dependent on autoantibody status. *Arthritis and Rheumatism,* 56**,** 2549-2556.

MARINOU, I., MAXWELL, J. R. & WILSON, A. G. 2010. Genetic influences modulating the radiological severity of rheumatoid arthritis. *Annals of the Rheumatic Diseases,* 69**,** 476-482.

MARINOU, I., TILL, S. H., MOORE, D. J. & WILSON, A. G. 2008. Lack of association or interactions between the IL-4, IL-4R alpha and IL-13 genes, and rheumatoid arthritis. *Arthritis Research & Therapy,* 10.

MARSH, S. G. E., ALBERT, E. D., BODMER, W. F., BONTROP, R. E., DUPONT, B., ERLICH, H. A., FERNANDEZ-VINA, M., GERAGHTY, D. E., HOLDSWORTH, R., HURLEY, C. K., LAU, M., LEE, K. W., MACH, B., MAIERS, M., MAYR, W. R., MUELLER, C. R., PARHAM, P., PETERSDORF, E. W., SASAZUKI, T., STROMINGER, J. L., SVEJGAARD, A., TERASAKI, P. I., TIERCY, J. M. & TROWSDALE, J. 2010. Nomenclature for factors of the HLA system, 2010. *Tissue Antigens,* 75**,** 291-455.

MASDOTTIR, B., JONSSON, T., MANFREOSDOTTIR, V., VIKINGSSON, A., BREKKAN, A. & VALDIMARSSON, H. 2000. Smoking, rheumatoid factor isotypes and severity of rheumatoid arthritis. *Rheumatology,* 39**,** 1202-1205.

MASSARDO, L., PONS-ESTEL, B. A., WOJDYLA, D., CARDIEL, M. H., GALARZA-MALDONADO, C. M., SACNUN, M. P., SORIANO, E. R., LAURINDO, I. M., ACEVEDO-VASQUEZ, E. M., CABALLERO-URIBE, C. V., PADILLA, O., GUIBERT-TOLEDANO, Z. M., DA MOTA, L. M., MONTUFAR, R. A., LINO-PEREZ, L., DIAZ-COTO, J., ACHURRA-CASTILLO, A. F., HERNANDEZ, J. A., ESTEVA-SPINETTI, M. H., RAMIREZ, L. A., PINEDA, C. & FURST, D. E. 2012. Early Rheumatoid Arthritis in Latin America: Low Socioeconomic Status Related to High Disease Activity at Baseline. *Arthritis Care & Research,* 64**,** 1135-1143.

MATTEY, D. L., HUTCHINSON, D., DAWES, P. T., NIXON, N. B., CLARKE, S., FISHER, J., BROWNFIELD, A., ALLDERSEA, J., FRYER, A. A. & STRANGE, R. C. 2002. Smoking and disease severity in rheumatoid arthritis - Association with polymorphism at the glutathione S-transferase M1 locus. *Arthritis and Rheumatism,* 46**,** 640-646.

MATTEY, D. L., NIXON, N., DAWES, P. T. & KERR, J. 2005. Association of polymorphism in the transforming growth factor beta 1 gene with disease outcome and mortality in rheumatoid arthritis. *Annals of the Rheumatic Diseases,* 64**,** 1190-1194.

MATTEY, D. L., NIXON, N. B., DAWES, P. T., OLLIER, W. E. R. & HAJEER, A. H. 2004. Association of matrix metalloproteinase 3 promoter genotype with disease outcome in rheumatoid arthritis. *Genes and Immunity,* 5**,** 147-149.

MAXWELL, J. R., GOWERS, I. R., MOORE, D. J. & WILSON, A. G. 2010. Alcohol consumption is inversely associated with risk and severity of rheumatoid arthritis. *Rheumatology,* 49**,** 2140-2146.

MEWAR, D., MARINOU, I., COOTE, A. L., MOORE, D. J., AKIL, M., SMILLIE, D., DICKSON, M. C., BINKS, M. H., MONTGOMERY, D. S. & WILSON, A. G. 2008. Association between radiographic severity of rheumatoid arthritis and shared epitope alleles: differing mechanisms of susceptibility and protection. *Annals of the Rheumatic Diseases,* 67**,** 980-983.

MEYER, J. M., HAN, J. F., SINGH, R. & MOXLEY, G. 1996. Sex influences on the penetrance of HLA shared-epitope genotypes for rheumatoid arthritis. *American Journal of Human Genetics,* 58**,** 371-383.

MEYER, P. W. A., HODKINSON, B., ALLY, M., MUSENGE, E., WADEE, A. A., FICKL, H., TIKLY, M. & ANDERSON, R. 2011. HLA-DRB1 shared epitope genotyping using the revised classification and its association with circulating autoantibodies, acute phase reactants, cytokines and clinical indices of disease activity in a cohort of South African rheumatoid arthritis patients. *Arthritis Research & Therapy,* 13.

MICHOU, L., CROISEAU, P., PETIT-TEIXEIRA, E., DU MONTCEL, S. T., LEMAIRE, I., PIERLOT, C., OSORIO, J., FRIGUI, W., LASBLEIZ, S., QUILLET, P., BARDIN, T., PRUM, B., CLERGET-DARPOUX, F., CORNELIS, F. & EUROPEAN CONSORTIUM, R. 2006. Validation of the reshaped shared epitope HLA-DRB1 classification in rheumatoid arthritis. *Arthritis Research & Therapy,* 8.

MIKULS, T. R., HUGHES, L. B., WESTFALL, A. O., HOLERS, V. M., PARRISH, L., VAN DER HEIJDE, D., VAN EVERDINGEN, M., ALARCON, G. S., CONN, D. L., JONAS, B., CALLAHAN, L. F., SMITH, E. A., GILKESON, G., HOWARD, G., MORELAND, L. W. & BRIDGES, S. L. 2008. Cigarette smoking, disease severity and autoantibody expression in African Americans with recent-onset rheumatoid arthritis. *Annals of the Rheumatic Diseases,* 67**,** 1529-1534.

MIN, J. Y., MIN, K. B., SUNG, J. & CHO, S. I. 2010. Linkage and Association Studies of Joint Morbidity from Rheumatoid Arthritis. *Journal of Rheumatology,* 37**,** 291-295.

MJAAVATTEN, M. D., VAN DER HEIJDE, D. M., UHLIG, T., HAUGEN, A. J., NYGAARD, H., BJORNEBOE, O. & KVIEN, T. K. 2011. Should Anti-citrullinated Protein Antibody and Rheumatoid Factor Status Be Reassessed During the First Year of Followup in Recent-Onset Arthritis? A Longitudinal Study. *Journal of Rheumatology,* 38**,** 2336-2341.

MORGAN, A. W., ROBINSON, J. I., CONAGHAN, P. G., MARTIN, S. G., HENSOR, E. M. A., MORGAN, M. D., STEINER, L., ERLICH, H. A., GOOI, H. C., BARTON, A., WORTHINGTON, J., EMERY, P., CONSORTIUM, U. & YEAR, C. 2010. Evaluation of the rheumatoid arthritis susceptibility loci HLA-DRB1, PTPN22, OLIG3/TNFAIP3, STAT4 and TRAF1/C5 in an inception cohort. *Arthritis Research & Therapy,* 12**,** R57.

MORGAN, A. W., THOMSON, W., MARTIN, S. G., CARTER, A. M., ERLICH, H. A., BARTON, A., HOCKING, L., REID, D. M., HARRISON, P., WORDSWORTH, P., STEER, S., WORTHINGTON, J., EMERY, P., WILSON, A. G., BARRETT, J. H. & YORKSHIRE EARLY ARTHRITIS REGISTER, U. K. R. A. G. C. 2009. Reevaluation of the Interaction Between HLA-DRB1 Shared Epitope Alleles, PTPN22, and Smoking in Determining Susceptibility to Autoantibody-Positive and Autoantibody-Negative Rheumatoid Arthritis in a Large UK Caucasian Population. *Arthritis and Rheumatism,* 60**,** 2565-2576.

NAM, E. J., KIM, K. H., HAN, S. W., CHO, C. M., LEE, J., PARK, J. Y. & KANG, Y. M. 2010. The-283C/T polymorphism of the DNMT3B gene influences the progression of joint destruction in rheumatoid arthritis. *Rheumatology International,* 30**,** 1299-1303.

NELL, V. P. K., MACHOLD, K. P., STAMM, T. A., EBERL, G., HEINZL, H., UFFMANN, M., SMOLEN, J. S. & STEINER, G. 2005. Autoantibody profiling as early diagnostic and prognostic tool for rheumatoid arthritis. *Annals of the Rheumatic Diseases,* 64**,** 1731-1736.

NEMEC, P., PAVKOVA-GOLDBERGOVA, M., STOURACOVA, M., VASKU, A., SOUCEK, M. & GATTEROVA, J. 2008. Polymorphism in the tumor necrosis factor-alpha gene promoter is associated with severity of rheumatoid arthritis in the Czech population. *Clinical Rheumatology,* 27**,** 59-65.

NICE. 2009. Rheumatoid arthritis: The management of rheumatoid arthritis in adults. NICE clinical guideline 79. [online] February 2009 (last updated 9th September 2010) National Collaborating Centre for Chronic Conditions Available from: http://guidance.nice.org.uk/CG79/Guidance/pdf/English. [Accessed 30 March 2011].

NICE. 2010. Adalimumab, etanercept, infliximab, rituximab and abatacept for the treatment of rheumatoid arthritis after the failure of a TNF inhibitor. NICE technology appraisal guidance 195. [online] March 2010. Available from: http://www.nice.org.uk/nicemedia/live/13108/50413/50413.pdf. [Accessed 30th March 2011].

NISHIMOTO, K., IKARI, K., MOCHIZUKI, T., TOMATSU, T., TOYAMA, Y., HARA, M., YAMANAKA, H., KAMATANI, N. & MOMOHARA, S. 2008. Lack of association between PADI4 and functional severity in Japanese rheumatoid arthritis patients. *Annals of the Rheumatic Diseases,* 67**,** 431-432.

NYAHLL-WAHLIN, B. M., PETERSSON, I. F., NILSSON, J. A., JACOBSSON, L. T. H., TURESSON, C. & GRP, B. S. 2009. High disease activity disability burden and smoking predict severe extra-articular manifestations in early rheumatoid arthritis. *Rheumatology,* 48**,** 416-420.

OEN, K., MALLESON, P. N., CABRAL, D. A., ROSENBERG, A. M., PETTY, R. E., NICKERSON, P. & REED, M. 2005. Cytokine genotypes correlate with pain and radiologically defined joint damage in patients with juvenile rheumatoid arthritis. *Rheumatology,* 44**,** 1115-1121.

OKADA, Y., TERAO, C., IKARI, K., KOCHI, Y., OHMURA, K., SUZUKI, A., KAWAGUCHI, T., STAHL, E. A., KURREEMAN, F. A. S., NISHIDA, N., OHMIYA, H., MYOUZEN, K., TAKAHASHI, M., SAWADA, T., NISHIOKA, Y., YUKIOKA, M., MATSUBARA, T., WAKITANI, S., TESHIMA, R., TOHMA, S., TAKASUGI, K., SHIMADA, K., MURASAWA, A., HONJO, S., MATSUO, K., TANAKA, H., TAJIMA, K., SUZUKI, T., IWAMOTO, T., KAWAMURA, Y., TANII, H., OKAZAKI, Y., SASAKI, T., GREGERSEN, P. K., PADYUKOV, L., WORTHINGTON, J., SIMINOVITCH, K. A., LATHROP, M., TANIGUCHI, A., TAKAHASHI, A., TOKUNAGA, K., KUBO, M., NAKAMURA, Y., KAMATANI, N., MIMORI, T., PLENGE, R. M., YAMANAKA, H., MOMOHARA, S., YAMADA, R., MATSUDA, F. & YAMAMOTO, K. 2012. Meta-analysis identifies nine new loci associated with rheumatoid arthritis in the Japanese population. *Nature Genetics,* 44**,** 511-+.

OROZCO, G., HINKS, A., EYRE, S., KE, X., GIBBONS, L. J., BOWES, J., FLYNN, E., MARTIN, P., WILSON, A. G., BAX, D. E., MORGAN, A. W., EMERY, P., STEER, S., HOCKING, L., REID, D. M., WORDSWORTH, P., HARRISON, P., THOMSON, W., BARTON, A., WORTHINGTON, J., WELLCOME TRUST CASE CONTROL, C. & CONSORTIUM, Y. 2009. Combined effects of three independent SNPs greatly increase the risk estimate for RA at 6q23. *Human Molecular Genetics,* 18**,** 2693-2699.

PALOMINO-MORALES, R., GONZALEZ-JUANATEY, C., VAZQUEZ-RODRIGUEZ, T. R., MIRANDA-FILLOY, J. A., LLORCA, J., MARTIN, J. & GONZALEZ-GAY, M. A. 2009. Interleukin-6 gene-174 promoter polymorphism is associated with endothelial dysfunction but not with disease susceptibility in patients with rheumatoid arthritis. *Clinical and Experimental Rheumatology,* 27**,** 964-970.

PAPADOPOULOS, N. G., ALAMANOS, Y., VOULGARL, P. V., EPAGELIS, E. K., TSIFETAKI, N. & DROSOS, A. A. 2005. Does cigarette smoking influence disease expression, activity and severity in early rheumatoid arthritis patients? *Clinical and Experimental Rheumatology,* 23**,** 861-866.

PARADOWSKA-GORYCKA, A., WOJTECKA-LUKASIK, E., TREFLER, J., WOJCIECHOWSKA, B., LACKI, J. K. & MASLINSKI, S. 2010. Association between IL-17F Gene Polymorphisms and Susceptibility to and Severity of Rheumatoid Arthritis (RA). *Scandinavian Journal of Immunology,* 72**,** 134-141.

PAWLIK, A., KURZAWSKI, A., FLORCZAK, A., SZKLARZ, B. G. & HERCZYNSKA, A. 2005a. IL1 beta+3953 exon 5 and IL-2-330 promoter polymorphisms in patients with rheumatoid arthritis. *Clinical and Experimental Rheumatology,* 23**,** 159-164.

PAWLIK, A., KURZAWSKI, M., SZKLARZ, B. G., HERCZYNSKA, M. & DROZDZIK, M. 2005b. Interleukin-10 promoter polymorphism in patients with rheumatoid arthritis. *Clinical Rheumatology,* 24**,** 480-484.

PAWLIK, A., WRZESNIEWSKA, J., FLORCZAK, M., GAWRONSKA-SZKLARZ, B. & HERCZYNSKA, M. 2005c. IL-6 promoter polymorphism in patients with rheumatoid arthritis. *Scandinavian Journal of Rheumatology,* 34**,** 109-113.

PAWLIK, A., WRZESNIEWSKA, J., FLORCZAK, M., GAWRONSKA-SZKLARZ, B. & HERCZYNSKA, M. 2005d. The -590 IL-4 promoter polymorphism in patients with rheumatoid arthritis. *Rheumatology International,* 26**,** 48-51.

PEDRESCHI, R., HERTOG, M. L. A. T. M., CARPENTIER, S. C., LAMMERTYN, J., ROBBEN, J., NOBEN, J.-P., PANIS, B., SWENNEN, R. & NICOLAI, B. M. 2008. Treatment of missing values for multivariate statistical analysis of gel-based proteomics data. *Proteomics,* 8**,** 1371-1383.

PIERER, M., KALTENHAUSER, S., ARNOLD, S., WAHLE, M., BAERWALD, C., HANTZCHEL, H. & WAGNER, U. 2006. Association of PTPN22 1858 single-nucleotide polymorphism with rheumatoid arthritis in a German cohort: higher frequency of the risk allele in male compared to female patients. *Arthritis Research & Therapy,* 8.

PIKWER, M., NILSSON, J.-A., BERGSTROM, U., JACOBSSON, L. T. H. & TURESSON, C. 2012. Early menopause and severity of rheumatoid arthritis in women older than 45 years. *Arthritis Research & Therapy,* 14.

PLANTINGA, T. S., FRANSEN, J., TAKAHASHI, N., STIENSTRA, R., VAN RIEL, P. L., VAN DEN BERG, W. B., NETEA, M. G. & JOOSTEN, L. A. B. 2010. Functional consequences of DECTIN-1 early stop codon polymorphism Y238X in rheumatoid arthritis. *Arthritis Research & Therapy,* 12.

POKORNY, V., MCQUEEN, F., YEOMAN, S., MERRIMAN, M., MERRIMAN, A., HARRISON, A., HIGHTON, J. & MCLEAN, L. 2005. Evidence for negative association of the chemokine receptor CCR5 d32 polymorphism with rheumatoid arthritis. *Annals of the Rheumatic Diseases,* 64**,** 487-490.

PROTS, I., SKAPENKO, A., WENDLER, J., MATTYASOVSZKY, S., YONE, C. L., SPRIEWALD, B., BURKHARDT, H., RAU, R., KALDEN, J. R., LIPSKY, P. E. & SCHULZE-KOOPS, H. 2006. Association of the IL4R single-nucleotide polymorphism I50V with rapidly erosive rheumatoid arthritis. *Arthritis and Rheumatism,* 54**,** 1491-1500.

PUGNER, K. M., SCOTT, D. I., HOLMES, J. W. & HIEKE, K. 2000. The costs of rheumatoid arthritis: An international long-term view. *Seminars in Arthritis and Rheumatism,* 29**,** 305-320.

PURCELL, S., NEALE, B., TODD-BROWN, K., THOMAS, L., FERREIRA, M. A. R., BENDER, D., MALLER, J., SKLAR, P., DE BAKKER, P. I. W., DALY, M. J. & SHAM, P. C. 2007. PLINK: A tool set for whole-genome association and population-based linkage analyses. *American Journal of Human Genetics,* 81**,** 559-575.

RAKOTOMALALA, R. 2005. TANAGRA: A free software for research and academic purposes. *Proceedings of EGC'2005 RNTI-E-3,* 2**,** 697–702.

RANNAR, S., GELADI, P., LINDGREN, F. & WOLD, S. 1995. A PLS KERNEL ALGORITHM FOR DATA SETS WITH MANY VARIABLES AND FEW OBJECTS .2. CROSS-VALIDATION, MISSING DATA AND EXAMPLES. *Journal of Chemometrics,* 9**,** 459-470.

RAYCHAUDHURI, S., THOMSON, B. P., REMMERS, E. F., EYRE, S., HINKS, A., GUIDUCCI, C., CATANESE, J. J., XIE, G., STAHL, E. A., CHEN, R., ALFREDSSON, L., AMOS, C. I., ARDLIE, K. G., BARTON, A., BOWES, J., BURTT, N. P., CHANG, M., COBLYN, J., COSTENBADER, K. H., CRISWELL, L. A., CRUSIUS, J. B. A., CUI, J., DE JAGER, P. L., DING, B., EMERY, P., FLYNN, E., HARRISON, P., HOCKING, L. J., HUIZINGA, T. W. J., KASTNER, D. L., KE, X., KURREEMAN, F. A. S., LEE, A. T., LIU, X., LI, Y., MARTIN, P., MORGAN, A. W., PADYUKOV, L., REID, D. M., SEIELSTAD, M., SELDIN, M. F., SHADICK, N. A., STEER, S., TAK, P. P., THOMSON, W., VAN DER HELM-VAN MIL, A. H. M., VAN DER HORST-BRUINSMA, I. E., WEINBLATT, M. E., WILSON, A. G., WOLBINK, G. J., WORDSWORTH, P., ALTSHULER, D., KARLSON, E. W., TOES, R. E. M., DE VRIES, N., BEGOVICH, A. B., SIMINOVITCH, K. A., WORTHINGTON, J., KLARESKOG, L., GREGERSEN, P. K., DALY, M. J., PLENGE, R. M., CONSORTIUM, B. & CONSORTIUM, Y. 2009. Genetic variants at CD28, PRDM1 and CD2/CD58 are associated with rheumatoid arthritis risk. *Nature Genetics,* 41**,** 1313-U76.

RIGBY, A. S., MACGREGOR, A. J. & THOMSON, G. 1998. HLA haplotype sharing in rheumatoid arthritis sibships: Risk estimates subdivided by proband genotype. *Genetic Epidemiology,* 15**,** 403-418.

RODRIGUEZ-RODRIGUEZ, L., LAMAS, J. R., VARADE, J., LOPEZ-ROMERO, P., TORNERO-ESTEBAN, P., ABASOLO, L., DE LA CONCHA, E. G., JOVER, J. A., URCELAY, E. & FERNANDEZ-GUTIERREZ, B. 2011. Plasma soluble IL-6 receptor concentration in rheumatoid arthritis: associations with the rs8192284 IL6R polymorphism and with disease activity. *Rheumatology International,* 31**,** 409-413.

RUIZ-ESQUIDE, V., GOMEZ-PUERTA, J. A., CANETE, J. D., GRAELL, E., VAZQUEZ, I., GUADALUPE ERCILLA, M., VINAS, O., GOMEZ-CENTENO, A., HARO, I. & SANMARTI, R. 2011. Effects of Smoking on Disease Activity and Radiographic Progression in Early Rheumatoid Arthritis. *Journal of Rheumatology,* 38**,** 2536-2539.

SAAG, K. G., CERHAN, J. R., KOLLURI, S., OHASHI, K., HUNNINGHAKE, G. W. & SCHWARTZ, D. A. 1997. Cigarette smoking and rheumatoid arthritis severity. *Annals of the Rheumatic Diseases,* 56**,** 463-469.

SALLIOT, C., DAWIDOWICZ, K., LUKAS, C., GUEDJ, M., PACCARD, C., BENESSIANO, J., DOUGADOS, M., NICAISE, P., MEYER, O. & DIEUDE, P. 2011. PTPN22 R620W genotype-phenotype correlation analysis and gene-environment interaction study in early rheumatoid arthritis: results from the ESPOIR cohort. *Rheumatology,* 50**,** 1802-1808.

SAS 2008. *SAS/STAT® 9.2 User's Guide.*, Cary, NC: SAS Institute Inc.

SCHAFER, J. L. & GRAHAM, J. W. 2002. Missing data: Our view of the state of the art. *Psychological Methods,* 7**,** 147-177.

SCHERER, H. U., VAN DER LINDEN, M. P. M., KURREEMAN, F. A. S., STOEKEN-RIJSBERGEN, G., LE CESSIE, S., HUIZINGA, T. W. J., VAN DER HELM-VAN MIL, A. H. & TOES, R. E. M. 2010. Association of the 6q23 region with the rate of joint destruction in rheumatoid arthritis. *Annals of the Rheumatic Diseases,* 69**,** 567-570.

SCOTT, D. L., WOLFE, F. & HUIZINGA, T. W. J. 2010. Rheumatoid arthritis. *Lancet,* 376**,** 1094-1108.

SCOTT, I. C., TAN, R., STAHL, D., STEER, S., LEWIS, C. M. & COPE, A. P. 2013. The protective effect of alcohol on developing rheumatoid arthritis: a systematic review and meta-analysis. *Rheumatology,* 52**,** 856-867.

SFAR, I., DHAOUADI, T., HABIBI, I., ABDELMOULA, L., MAKHLOUF, M., BEN ROMDHANE, T., JENDOUBI-AYED, S., AOUADI, H., BEN ABDALLAH, T., AYED, K., ZOUARI, R. & LAKHOUA-GORGI, Y. 2009. Functional polymorphisms of PTPN22 and FcgR genes in Tunisian patients with rheumatoid arthritis. *Arch Inst Pasteur Tunis,* 86**,** 51-62.

SHEEDY, F. J., MARINOU, I., O'NEILL, L. A. J. & WILSON, A. G. 2008. The Mal/TIRAP S180L and TLR4 G299D polymorphisms are not associated with susceptibility to, or severity of, rheumatoid arthritis. *Annals of the Rheumatic Diseases,* 67**,** 1328-1331.

SIGNORINI, D. F. 1991. SAMPLE-SIZE FOR POISSON REGRESSION. *Biometrika,* 78**,** 446-450.

SJOSTROM, M., WOLD, S., LINDBERG, W., PERSSON, J. A. & MARTENS, H. 1983. A MULTIVARIATE CALIBRATION-PROBLEM IN ANALYTICAL-CHEMISTRY SOLVED BY PARTIAL LEAST-SQUARES MODELS IN LATENT-VARIABLES. *Analytica Chimica Acta,* 150**,** 61-70.

SKOV, T., BALLABIO, D. & BRO, R. 2008. Multiblock variance partitioning: A new approach for comparing variation in multiple data blocks. *Analytica Chimica Acta,* 615**,** 18-29.

SODERLIN, M. K., PETERSSON, I. F., BERGMAN, S., SVENSSON, B. & GRP, B. S. 2011. Smoking at onset of rheumatoid arthritis (RA) and its effect on disease activity and functional status: experiences from BARFOT, a long-term observational study on early RA. *Scandinavian Journal of Rheumatology,* 40**,** 249-255.

STAHL, E. A., RAYCHAUDHURI, S., REMMERS, E. F., XIE, G., EYRE, S., THOMSON, B. P., LI, Y., KURREEMAN, F. A. S., ZHERNAKOVA, A., HINKS, A., GUIDUCCI, C., CHEN, R., ALFREDSSON, L., AMOS, C. I., ARDLIE, K. G., BARTON, A., BOWES, J., BROUWER, E., BURTT, N. P., CATANESE, J. J., COBLYN, J., COENEN, M. J. H., COSTENBADER, K. H., CRISWELL, L. A., CRUSIUS, J. B. A., CUI, J., DE BAKKER, P. I. W., DE JAGER, P. L., DING, B., EMERY, P., FLYNN, E., HARRISON, P., HOCKING, L. J., HUIZINGA, T. W. J., KASTNER, D. L., KE, X., LEE, A. T., LIU, X., MARTIN, P., MORGAN, A. W., PADYUKOV, L., POSTHUMUS, M. D., RADSTAKE, T. R. D. J., REID, D. M., SEIELSTAD, M., SELDIN, M. F., SHADICK, N. A., STEER, S., TAK, P. P., THOMSON, W., VAN DER HELM-VAN MIL, A. H. M., VAN DER HORST-BRUINSMA, I. E., VAN DER SCHOOT, C. E., VAN RIEL, P. L. C. M., WEINBLATT, M. E., WILSON, A. G., WOLBINK, G. J., WORDSWORTH, B. P., WIJMENGA, C., KARLSON, E. W., TOES, R. E. M., DE VRIES, N., BEGOVICH, A. B., WORTHINGTON, J., SIMINOVITCH, K. A., GREGERSEN, P. K., KLARESKOG, L., PLENGE, R. M., CONSORTIUM, B. & CONSORTIUM, Y. 2010. Genome-wide association study meta-analysis identifies seven new rheumatoid arthritis risk loci. *Nature Genetics,* 42**,** 508-U56.

STEER, S., LAD, B., GRUMLEY, J. A., KINGSLEY, G. H. & FISHER, S. A. 2005. Association of R602W in a protein tyrosine phosphatase gene with a high risk of rheumatoid arthritis in a British population: evidence for an early onset/disease severity effect. *Arthritis and Rheumatism,* 52**,** 358-360.

SUZUKI, T., IKARI, K., YANO, K., INOUE, E., TOYAMA, Y., TANIGUCHI, A., YAMANAKA, H. & MOMOHARA, S. 2013. PADI4 and HLA-DRB1 Are Genetic Risks for Radiographic Progression in RA Patients, Independent of ACPA Status: Results from the IORRA Cohort Study. *Plos One,* 8.

SUZUKI, T., TSUTSUMI, A., SUZUKI, H., SUZUKI, E., SUGIHARA, M., MURAKI, Y., HAYASHI, T., CHINO, Y., GOTO, D., MATSUMOTO, I., ITO, S., MIYAZAWA, K. & SUMIDA, T. 2008. Tristetraprolin (TTP) gene polymorphisms in patients with rheumatoid arthritis and healthy individuals. *Modern Rheumatology,* 18**,** 472-479.

SYMMONS, D., TURNER, G., WEBB, R., ASTEN, P., BARRETT, E., LUNT, M., SCOTT, D. & SILMAN, A. 2002. The prevalence of rheumatoid arthritis in the United Kingdom: new estimates for a new century. *Rheumatology,* 41**,** 793-800.

SYMMONS, D. P. M., BARRETT, E. M., BANKHEAD, C. R., SCOTT, D. G. I. & SILMAN, A. J. 1994. THE INCIDENCE OF RHEUMATOID-ARTHRITIS IN THE UNITED-KINGDOM - RESULTS FROM THE NORFOLK ARTHRITIS REGISTER. *British Journal of Rheumatology,* 33**,** 735-739.

TAYLOR, L. H., TWIGG, S., WORTHINGTON, J., EMERY, P., MORGAN, A. W., WILSON, A. G. & TEARE, M. D. 2013. Metaanalysis of the Association of Smoking and PTPN22 R620W Genotype on Autoantibody Status and Radiological Erosions in Rheumatoid Arthritis. *J Rheumatol,* 40**,** 1048-53.

TEARE, M. D., KNEVEL, R., MORGAN, M. D., KLESZCZ, A., EMERY, P., MOORE, D. J., CONAGHAN, P., HUIZINGA, T. W., MORGAN, A. W., VAN DER HELM-VAN MIL, A. H. & WILSON, A. G. 2013. Allele dose association of the C5orf30 rs26232 variant with joint damage in rheumatoid arthritis. *Arthritis Rheum*.

TEH, C. L. & WONG, J. S. 2011. Impact of tight control strategy on rheumatoid arthritis in Sarawak. *Clinical Rheumatology,* 30**,** 615-621.

TEZENAS DU MONTCEL, S., MICHOU, L., PETIT-TEIXEIRA, E., OSORIO, J., LEMAIRE, I., LASBLEIZ, S., PIERLOT, U., QUILLET, P., BARDIN, T., PRUM, B., CORNELIS, F. O. & CLERGET-DARPOUX, F. 2005. New classification of HLA-DRB1 alleles supports the shared epitope hypothesis of rheumatoid arthritis susceptibility. *Arthritis and Rheumatism,* 52**,** 1063-1068.

TEZENAS DU MONTCEL, S., REVIRON, D., GENIN, E., ROUDIER, J., MERCIER, P. & CLERGET-DARPOUX, F. 2000. Modeling the HLA component in rheumatoid arthritis: Sensitivity to DRB1 allele frequencies. *Genetic Epidemiology,* 19**,** 422-428.

THOMSON, W., BARTON, A., KE, X., EYRE, S., HINKS, A., BOWES, J., DONN, R., SYMMONS, D., HIDER, S., BRUCE, I. N., WILSON, A. G., MARINOU, I., MORGAN, A., EMERY, P., CARTER, A., STEER, S., HOCKING, L., REID, D. M., WORDSWORTH, P., HARRISON, P., STRACHAN, D., WORTHINGTON, J., CONSORTIUM, W. & CONSORTIUM, Y. 2007. Rheumatoid arthritis association at 6q23. *Nature Genetics,* 39**,** 1431-1433.

THYBERG, I., HASS, U. A. M., NORDENSKIOLD, U., GERDLE, B. & SKOGH, T. 2005. Activity limitation in rheumatoid arthritis correlates with reduced grip force regardless of sex: The Swedish TIRA Project. *Arthritis & Rheumatism-Arthritis Care & Research,* 53**,** 886-896.

TOONEN, E. J. M., COENEN, M. J. H., KIEVIT, W., FRANSEN, J., EIJSBOUTS, A. M., SCHEFFER, H., RADSTAKE, T., CREEMERS, M. C. W., DE ROOIJ, D., VAN RIEL, P., FRANKE, B. & BARRERA, P. 2008. The tumour necrosis factor receptor superfamily member 1b 676T > G polymorphism in relation to response to infliximab and adalimumab treatment and disease severity in rheumatoid arthritis. *Annals of the Rheumatic Diseases,* 67**,** 1174-1177.

TRYGG, J. & WOLD, S. 2002. Orthogonal projections to latent structures (O-PLS). *Journal of Chemometrics,* 16**,** 119-128.

TURKMEN, A. & LIN, S. 2012. An Optimum Projection and Noise Reduction Approach for Detecting Rare and Common Variants Associated with Complex Diseases. *Human Heredity,* 74**,** 51-60.

VAJARGAH, K. F., SADEGHI-BAZARGANI, H., MEHDIZADEH-ESFANJANI, R., SAVADI-OSKOUEI, D. & FARHOUDI, M. 2012. OPLS statistical model versus linear regression to assess sonographic predictors of stroke prognosis. *Neuropsychiatric Disease and Treatment,* 8**,** 387-392.

VALLE, Y., PADILLA-GUTIERREZ, J. R., TORRES-CARRILLO, N. M., LEDEZMA-LOZANO, I. Y., CORONA-SANCHEZ, E. G., VAZQUEZ-DEL MERCADO, M., RANGEL-VILLALOBOS, H., GAMEZ-NAVA, J. I., GONZALEZ-LOPEZ, L. & MUNOZ-VALLE, J. F. 2010. The-383A > C TNFRI polymorphism is associated with soluble levels and clinical activity in rheumatoid arthritis. *Rheumatology International,* 30**,** 655-659.

VAN AKEN, J., VAN BILSEN, J. H. M., ALLAART, C. F., HUIZINGA, T. W. J. & BREEDVELD, F. C. 2003. The Leiden Early Arthritis Clinic. *Clinical and Experimental Rheumatology,* 21**,** S100-S105.

VAN DE GEIJN, F. E., HAZES, J. M. W., GELEIJNS, K., EMONTS, M., JACOBS, B. C., DUFOUR-VAN DEN GOORBERGH, B. C. M. & DOLHAIN, R. 2008. Mannose-binding lectin polymorphisms are not associated with rheumatoid arthritis - confirmation in two large cohorts. *Rheumatology,* 47**,** 1168-1171.

VAN DER HELM-VAN MIL, A. H. M. & HUIZINGA, T. W. J. 2008. Advances in the genetics of rheumatoid arthritis point to subclassification into distinct disease subsets. *Arthritis Research & Therapy,* 10.

VAN DER LINDEN, M. P. M., FEITSMA, A. L., LE CESSIE, S., KERN, M., OLSSON, L. M., RAYCHAUDHURI, S., BEGOVICH, A. B., CHANG, M., CATANESE, J. J., KURREEMAN, F. A. S., VAN NIES, J., VAN DER HEIJDE, D. M., GREGERSEN, P. K., HUIZINGA, T. W. J., TOES, R. E. M. & VAN DER HELM-VAN MIL, A. H. M. 2009. Association of a Single-Nucleotide Polymorphism in CD40 With the Rate of Joint Destruction in Rheumatoid Arthritis. *Arthritis and Rheumatism,* 60**,** 2242-2247.

VAN DER VOET, H. 1994. Comparing the Predictive Accuracy of Models Using a Simple Randomization Test. *Chemometrics and Intelligent Laboratory Systems,* 25**,** 313-323.

VAN DER WOUDE, D., HOUWING-DUISTERMAAT, J. J., TOES, R. E. M., HUIZINGA, T. W. J., THOMSON, W., WORTHINGTON, J., VAN DER HELM-VAN MIL, A. H. M. & DE VRIES, R. R. P. 2009. Quantitative Heritability of Anti-Citrullinated Protein Antibody-Positive and Anti-Citrullinated Protein Antibody-Negative Rheumatoid Arthritis. *Arthritis and Rheumatism,* 60**,** 916-923.

VAN GAALEN, F. A., VAN AKEN, J., HUIZINGA, T. W. J., SCHREUDER, G. M. T., BREEDVELD, F. C., ZANELLI, E., VAN VENROOIJ, W. J., VERWEIJ, C. L., TOES, R. E. M. & DE VRIES, R. R. P. 2004. Association between HLA class II genes and autoantibodies to cyclic citrullinated peptides (CCPs) influences the severity of rheumatoid arthritis. *Arthritis and Rheumatism,* 50**,** 2113-2121.

VANZEBEN, D., HAZES, J. M. W., VANDENBROUCKE, J. P., DIJKMANS, B. A. C. & CATS, A. 1990. DIMINISHED INCIDENCE OF SEVERE RHEUMATOID-ARTHRITIS ASSOCIATED WITH ORAL-CONTRACEPTIVE USE. *Arthritis and Rheumatism,* 33**,** 1462-1465.

VARMA, S. & SIMON, R. 2006. Bias in error estimation when using cross-validation for model selection. *Bmc Bioinformatics,* 7.

VELPULA, U. D., AGRAWAL, S., THOMAS, J., PRABU, V. N. N., RAJASEKHAR, L. & NARSIMULU, G. 2011. Low Body Mass Index Is Adversely Associated with Radiographic Joint Damage in Indian Patients with Early Rheumatoid Arthritis. *Journal of Rheumatology,* 38**,** 434-438.

VINZI, V. E., CHIN, W., HENSELER, J. & WANG, H. 2010. *Handbook of Partial Least Squares: Concepts, Methods and Applications*, Springer-Verlag Berlin Heidelberg.

WAGENER, F., TOONEN, E. J. M., WIGMAN, L., FRANSEN, J., CREEMERS, M. C. W., RADSTAKE, T., COENEN, M. J. H., BARRERA, P., VAN RIEL, P. & RUSSEL, F. G. M. 2008. HMOX1 Promoter Polymorphism Modulates the Relationship Between Disease Activity and Joint Damage in Rheumatoid Arthritis. *Arthritis and Rheumatism,* 58**,** 3388-3393.

WAGNER, U., KALTENHAUSER, S., PIERER, M., SEIDEL, W., TROLTZSCH, M., HANTZSCHEL, H., KALDEN, J. R. & WASSMUTH, R. 2003. Prospective analysis of the impact of HLA-DR and -DQ on joint destruction in recent-onset rheumatoid arthritis. *Rheumatology,* 42**,** 553-562.

WANG, T., HO, G., YE, K., STRICKLER, H. & ELSTON, R. C. 2009. A Partial Least-Square Approach for Modeling Gene-Gene and Gene-Environment Interactions When Multiple Markers Are Genotyped. *Genetic Epidemiology,* 33.

WARE, J. E., KOSINSKI, M. & GANDEK, B. 2000. *SF-36 ® Health Survey: Manual & Interpretation Guide*, Lincoln, RI: QualityMetric Incorporated.

WEHRENS, R., FRANCESCHI, P., VRHOVSEK, U. & MATTIVI, F. 2011. Stability-based biomarker selection. *Analytica Chimica Acta,* 705**,** 15-23.

WESTHOFF, G., RAU, R. & ZINK, A. 2008. Rheumatoid arthritis patients who smoke have a higher need for DMARDs and feel worse, but they do not have more joint damage than non-smokers of the same serological group. *Rheumatology,* 47**,** 849-854.

WINHAM, S. J., COLBY, C. L., FREIMUTH, R. R., WANG, X., DE ANDRADE, M., HUEBNER, M. & BIERNACKA, J. M. 2012. SNP interaction detection with Random Forests in high-dimensional genetic data. *Bmc Bioinformatics,* 13.

WOLD, H. 1975. Path models with latent variables: the NIPALS approach. *Quantitative sociology: International perspectives on mathematical and statistical modeling.***,** pp. 307–357.

WOLD, H. 1985. Partial least squares, pp. 581–591 in Samuel Kotz and Norman L. Johnson, eds., *Encyclopedia of statistical sciences*, Vol. 6, New York: Wiley, 1985.

YUN, H. R., LEE, S. O., CHOI, E. J., SHIN, H. D., JUN, J. B. & BAE, S. C. 2008. Cyclooxygenase-2 polymorphisms and risk of rheumatoid arthritis in Koreans. *Journal of Rheumatology,* 35**,** 763-769.

ZAPICO, I., COTO, E., RODRIGUEZ, A., ALVAREZ, C., TORRE, J. C. & ALVAREZ, V. 2000a. A DNA polymorphism at the alpha(2)-macroglobulin gene is associated with the severity of rheumatoid arthritis. *Journal of Rheumatology,* 27**,** 2308-2311.

ZAPICO, I., COTO, E., RODRIGUEZ, A., ALVAREZ, C., TORRE, J. C. & ALVAREZ, V. 2000b. CCR5 (chemokine receptor-5) DNA-polymorphism influences the severity of rheumatoid arthritis. *Genes and Immunity,* 1**,** 288-289.

Appendix A: Ethics approval for secondary analysis

The University Of Sheffield.

Cheryl Oliver
Ethics Committee Administrator

Regent Court
30 Regent Street
Sheffield S1 4DA

Telephone: +44 (0) 114 2220871
Fax: +44 (0) 114 272 4095 (non confidential)
Email: c.a.oliver@sheffield.ac.uk

19 January 2011

Lyn Taylor
ScHARR

Dear Lyn

Statistical modelling of markers of severity in Rheumatoid Arthritis

I am pleased to inform you your supervisor has reviewed your project and classed it as 'low risk' so you can proceed with your research. The research must be conducted within the requirements of the hosting/employing organisation or the organisation where the research is being undertaken.

I have received a hard copy of your student declaration together with your Supervisor's confirmation for research that does not involve human participants and that you will be undertaking research which involves analysis of already existing data ('secondary data').

Yours sincerely

Cheryl Oliver
Ethics Committee Administrator

Cc: Dawn Teare

Appendix B: Table of evidence of gene association with severity of rheumatoid arthritis

Acronyms: ACPA=Anti-Citrullinated protein antibody, CRP= C-Reactive protein, CI=Confidence interval, DAS=disease activity score, DMARD=Disease modifying anti-rheumatoid drug, (M)HAQ=(Modified) health assessment questionnaire, MTX=Methotrexate, OR=Odds ratio, RA=Rheumatoid arthritis, RAAD= Rheumatoid Arthritis Articular Damage score, RF=Rheumatoid factor, SJC=Swollen joint count, TJC=Tender joint count, UNK=unknown information, VAS=Visual analogue scale, VNTR=Variable number of tandem repeats.

Gene/ rs number/ Severity marker/ Reference	Sample size	Evidence of association	No evidence of association	Analysis performed
α2m/ UNK/ Disease activity/ (Zapico et al., 2000a)	71 severe RA and 89 non-severe RA	There was a significantly higher frequency of carriers with the α2m deletion allele in patients with an early active severe RA, compared to non-severe RA (p = 0.037), in patients with severe RA compared to non-severe RA (p = 0.017) and in patients with 5 or more episodes of acute exacerbation of disease activity per year (n = 39) compared to those with none (n = 46) (p = 0.002)		UNK
AMPD1/ rs17602729/ DAS28/ (Grabar et al., 2010)	211 RA	Carriers of AMPD1 34T (rs17602729) allele had a 3.8-fold higher probability for lower DAS28 (score <=3.2) compared with non-carriers (p=0.012, OR=3.786, 95% CI 1.347 to 10.642). After exclusion of patients co-treated with other DMARDs, the effect was even stronger (p=0.006, OR=6.729, 95% CI 1.741 to 26.007)		Binary logistic regression modelling the percentage with low disease activity (DAS28 <=3.2) vs. moderate/high (DAS28 >3.2) corrected for gender, MTX treatment duration, MTX dose and presence of RF and ACPA antibodies
C5orf30/ rs26232/ Larsen and Sharp score/ (Teare et al., 2013)	1884 RA	Reduction in joint damage with the T allele. Found in 885 cohort, median Larsen score CC=31, CT=27, TT=16, $p=4x10^{-4}$. Meta-analysis replication in 2 other cohorts (N=581and 418) found severity OR of 0.90 (0.84-0.96, p=0.004).		Fixed effects meta-analysis.
Calprotectin/ None (cell protein mg/l)/ Sharp, RAAD/	145 RA	Calprotectin showed a highly significant correlation with modified Sharp score (r = 0.43, p<0.001) and RAAD (r = 0.40, p<0.001) which was		Spearman's rank correlation. Linear multiple regression analyses with the modified Sharp

Gene/ rs number/ Severity marker/ Reference	Sample size	Evidence of association	No evidence of association	Analysis performed
(Hammer et al., 2007)		maintained after adjustment for CRP, ESR, RF, DAS28, sex, and age in a multiple regression analysis. The parameter estimates equated to a Sharp score increase of 5.49 (SE=2.30), p = 0.018 for every 1 point increase in calprotectin and a RAAD increase of 1.12 (SE=0.55), p = 0.04 for every one point increase in calprotectin		score and RAAD score as dependent variables and calprotectin, CRP, ESR, RF, DAS28, sex, and age as independent variables.
CARD8 (TUCAN)/ rs2043211/ Larsen, DAS28, SJC, TJC, ESR/ (Kastbom et al., 2010)	560 RA	Patients carrying CARD8-X had significantly higher DAS28 (p=0.02), ESR (p=0.004) and TJC (0.02) than those carrying CARD8-CC over the 24 months. No effect was seen for SJC or Larsen score at baseline or 24 months		Analysis of variance (ANOVA) for repeated measurements was used to compare data collected at several time points.
CARD8 (TUCAN) and CIAS1/ rs2043211 and UNK/ DAS28, ESR, physicians global assessment, CRP/ (Kastbom et al., 2008)	174 RA	Over a three year follow-up, patients with presence of at least one variant allele in both genes (CIAS1/TUCAN -/-) showed significantly higher disease activity at most time points. In patients presenting CIAS1/TUCAN +/+, only one (2%) received TNF-blocking therapy compared to seven (37%) in the CIAS1/TUCAN -/- group		Mann-Whitney U test
CIAS1/ rs35829419/ Larsen, DAS28, SJC, TJC, ESR/ (Kastbom et al., 2010)	560 RA		There were no associations between NLRP3 (previously called CIAS1) and disease activity measures	ANOVA for repeated measurements was used to compare data collected at several time points.
CCR5/ 32-bit deletion/ Erosions, ESR/CRP, anaemia/ (Zapico et al., 2000b)	160 RA	Carriers of the CCR5- Δ32 allele were at a significantly higher frequency (P = 0.012) in non-severe RA compared to early severe RA patients (17% vs. 4%).		Yates chi-square test to compare patients with early severe RA vs. non-severe.
CCR5/ 32-bit deletion/ Baseline or follow up TJC, SJC, RF, ESR, CRP, DMARD usage,	92 RA		After correcting for multiple testing, CCR5- Δ32 status was not associated with a significant difference in	Mann-Whitney U test

Gene/ rs number/ Severity marker/ Reference	Sample size	Evidence of association	No evidence of association	Analysis performed
HAQ, radiographic erosion status/ (Pokorny et al., 2005)			disease severity	
CCR5/ 32-bit deletion/ UNK/ (Graudal, 2004)	682 RA		The CCR5-Δ32 was not associated with disease severity	UNK
CCR5/ -1118 (ins/del) rs10577983/ Sharp score/ (Han et al., 2012b)	357	Significant increase in total Sharp score associated with both the -1118 CTAT (insertion/deletion) in CCR5 (rs10577983) (p=0.048) and 303 A>G (rs1799987) (p=0.048). In addition, when analysing the erosion score alone, there was an increase in the statistical significance (corrected for multiple testing p values of p=0.028 and p=0.028 respectively).		UNK
CD40/ rs4810485/ Sharp/van der Heijde scores/ (van der Linden et al., 2009)	563 RA + 383 RA in rep-lication	The TT and GT/GG genotype of CD40 (rs4810485) was associated with a higher rate of joint destruction in ACPA positive RA patients (back transformed regression coefficient of 1.12 times greater increase in the Sharp score per year (95% CI=1.04-1.21) p=0.003). This finding was not statistical significant after Bonferroni correction. In replication using a perfect proxy for rs4810485, a higher progression rate of 3.40 sharp units/year in the TT genotype compared to 2.83 and 1.83 in the GT and GG genotypes respectively was observed (p=0.021)		Sharp scores were presented as medians over time and were log-transformed in the analysis. A linear model for longitudinal data was used to compare progression rates between groups. Bonferroni adjustment P<0.008 (6 SNPs)
CDK6 / rs42041/ Sharp/van der Heijde scores/ (van der Linden et al., 2009)	563 RA + 383 RA in rep-lication	The GG and CC/CG genotype of CDK6 (rs42041) is associated with a higher rate of joint destruction in ACPA positive RA patients (back transformed regression coefficient of 1.09 times greater Increase in the Sharp score per year (95%		Sharp scores were presented as medians over time and were log-transformed in the analysis. A linear model for longitudinal data was used to compare progression

Gene/ rs number/ Severity marker/ Reference	Sample size	Evidence of association	No evidence of association	Analysis performed
		CI=1.02-1.16) p=0.012. This finding was not statistical significant after Bonferroni adjustment and significance could not be replicated using imputed data for rs42041 (2.76 sharp units/year in the GG genotype compared to 2.38 and 2.07 in the CG and CC genotypes respectively (p=0.327) although there is still evidence of a linear trend.		rates between groups. Bonferroni adjustment P<0.008 (6 SNPs)
Chondromodulin-II (ChM-II)/ UNK/ Modified Larsen score/ (Graessler et al., 2005)	204 RA	Presence of the ChM-II 172 A allele increases x-ray damage independent of SE. Larsen scores were significantly higher in RA patients carrying the 172AA genotype (Larsen score = 96.8), than in RA patients with the 172GA (Larsen score = 69.5) or 172GG (Larsen score = 54.8; p = 0.001) genotypes. ORs to develop more severe radiographic joint damage (Larsen score > 90; above 75th percentile) were 4 and 15.5 for the 172GA and 172AA genotypes, respectively. Presence of a 172A allele increased the risk for enhanced radiographic damage 3-fold.		Cochran Armitage trend test
Cyclooxygenase-2 (COX-2)/ -765G→C/ Anatomical stage according to Steinbrocker/ (Lee et al., 2006)	258 RA	No association was observed between COX-2 genotype and severity of RA. However, among those without the shared epitope (SE), carriers of the low activity C allele had a lower risk of RA and less severe form of RA than subjects with the G/G genotype. The OR (95% CI) was 0.04 (0.01-0.41) for severity of RA		OR and 95% CIs
Cyclo-oxygenase-2 (COX-2)/ 23 SNP inc -1329A→G, -899G→C SNP and 6365T→C/ Radiologic severity/ (Yun et al., 2008)	1201 RA		Radiologic severity of RA was not associated with COX-2 polymorphisms	Logistic regression models were used to calculate OR and 95% CIs
Cytochrome (CYP) CYP1A2/	1268 RA	A marked reduction in disease severity		ORs and 95% CIs adjusted for

Gene/ rs number/ Severity marker/ Reference	Sample size	Evidence of association	No evidence of association	Analysis performed
rs762551/ Anatomical stage according to Steinbrocker/ (Cornelis et al., 2010)		associated with the CYP1A2 C allele (-163 A→C rs762551) was found among cases homozygous for the SE. OR (95% CI): 0.30 (0.09-1.01), P=0.05		age, sex, duration of disease and duration of treatment.
DNA Methyltrasferase (DNMT) 3B/ -283 C→T/ Modified Sharp score/ (Nam et al., 2010)	309 RA	The - 283 C/T polymorphism (Chr 20q 11.2) of the DNMT3B gene contributes to the progression of joint destruction in RA. In the carriers of CT/TT, the slope of the regression line was significantly steeper than in the carriers of the CC genotype ($y=9.546x +19.998$, $r^2=0.810$, vs. $y=6.185x+34.424$, $r^2=0.536$; p=0.014)		Plotted modified sharp score against disease duration. Fitted regression lines for each genotype (CT+TT vs. CC). Differences in the slopes of the regression lines were analysed using an interaction between a dummy and time variable based on a multiple linear regression model.
Endothelial nitric oxide synthase (eNOS)/ T-786C/ Extra-articular manifestations/ (Brenol et al., 2009)	105 RA	The C/C genotype carriers were more likely to demonstrated extra-articular manifestations compared with the heterozygous and T/T homozygous taken genotypes taken together (OR = 4.9, 95% CI = 1.3-18.9, P = 0.022)		Chi-square for categorical variables, t-test/ANOVA for normally distributed variables and if assumptions are not met fitted a Mann-whitney U test or Kruskal wallis test.
FcgR Receptors for the Fc fragment of IgG/ FcgRIIIa-V/V158/ Joint erosions/ (Sfar et al., 2009)	133 RA	FcgRIIIa-V/V158 was the most important FcgR genotype for the severe disease subset with joint erosions. Patients with FcgRIIIb-NA2/NA2 genotype had an earlier incidence of clinical symptoms		UNK
FCRL-3/ -169 T>C, rs7528684/ Sharp score/ (Han et al., 2012a)	227 RA	Higher Sharp scores associated with the CC genotype for patients in the ≥10 year disease duration subgroup (p=0.034)		UNK
FCRL-3/	652 RA	The CC genotype was associated with 10 year		Multivariate linear and logistic

Gene/ rs number/ Severity marker/ Reference	Sample size	Evidence of association	No evidence of association	Analysis performed
-169 T>C, rs7528684/ Radiographic progression/ (Maehlen et al., 2011)		radiographic progression.		regression analyses after adjustment for ACPA, ESR age, and sex
FCRL-3/ -169 T>C, rs7528684/ Destructive vs non/ (Chen et al., 2011)	670 RA	There was an increase in CC+CT genotypes for patients with destructive disease compared to non-destructive disease (OR=1.672, 95% CI: 1.149-2.432, p=0.007)		UNK
FCRL-3/ -169 T>C, rs7528684/ Erosions/ (Bajpai et al., 2012)	58 RA	FCRL-3 -169C allele was overrepresented in patients with erosive RA.		UNK
Glutathione S-transferases (GST)/ /GSTM1/ Larsen score & HAQ/ (Mattey et al., 2002)	164 RA	Disease outcome in female RA patients with a history of smoking was significantly worse than in those who had never smoked. Smoking was significantly associated with the most severe disease (higher Larsen score) in patients who carried the GSTM1-null polymorphism. This association may be due in part to a relationship between the GSTM1 polymorphism and RF production in smokers		Multiple regression analyses, with correction for age and disease duration
Glutathione S-transferases (GST)/ /GSTM1, GSTT1, GSTP1/ DAS28/ (Bohanec Grabar et al., 2009)	213 RA	Patients with GSTT1-null polymorphism (deletion) had a higher risk for developing high activity RA than the patients with GSTT1 genes present (p=0.028, OR=2.761, 95% CI=1.114-6.843). In the group of smokers, patients with GSTT1 deletion had an 8.5-fold higher risk for developing high disease activity than patients without the deletion (p=0.004, OR=8.640, 95% CI=1.995-37.426)	GSTM1 –null polymorphism (deletion) and GSTP1 polymorphisms were not associated with the disease activity	
HLA-DRB1/ DRB1*01 and DRB1*04/ Larsen scores/	87 RA	*01 or *04 are significantly associated with higher x-ray damage		HLA markers were evaluated by univariate comparisons and by multiple logistic regression of

Gene/ rs number/ Severity marker/ Reference	Sample size	Evidence of association	No evidence of association	Analysis performed
(Wagner et al., 2003)				progression over time.
HLA-DRB1 / DRB1*04/ DAS, Larsen score/ (Farouk et al., 2009)	29 RA	HLA-DRB1*04 alleles (n=12) were significantly expressed among RA patients. (N=9, 75% with active disease vs. N=3, 25% with inactive disease, p<0.05 and N=3, 25% with non-erosive disease vs. N=9, 75% with erosive disease, p<0.05)		Students T-test of proportion of patients
HLA-DRB1/ S2, S3D / Sharp-van der Heijde method/ (Gourraud et al., 2006)	144	The presence of S_2 alleles (HLA-DRB1*0401 and HLA-DRB1*1303) were associated with severe forms of RA (P = 0.004). A significant dose effect was observed (P = 0.01). The presence of S3D alleles (HLA-DRB1*11001, HLA-DRB1*1104, HLA-DRB1*12 and HLA-DRB1*16) were associated with benign forms of RA (P<0.0001) and a significant dose effect was observed (P<0.01)		Non-parametric statistical tests (no further information available).
HLA-DRB1/ S2, S1/ Larsen score/ (Mewar et al., 2008)	962	S_2 alleles (KRAA) coding for (HLA-DRB1*0401 and HLA-DRB1*1303) were associated with more severe structural damage (Kruskal wallis test for 0 vs. 1 vs. 2, S_2 alleles, Larsen score =26, 29 and 41 respectively, p=0.0059 and Mann-Whitney U test for S2/S2 genotype vs. X/X genotype, Larsen score=41 and 22 respectively, p=0.01). S_2 alleles were also found to affect the proportion of patients being RF positive (0 vs. 1 vs. 2, S_2 alleles having 62%, 75% and 80% RF positive patients respectively, p=<0.001) and the proportion of ACPA positive patients (0 vs. 1 vs. 2, S_2 alleles having 66%, 86% and 91% positive respectively, p=<0.001) S_1 alleles were associated with less severe disease (0 vs. 1 vs. 2, S_1 alleles, Larsen score =28, 20, 18 respectively, p=0.011 and S1/S1 genotype vs. X/X genotype, Larsen score=18 and 22		Kruskal wallis test comparing alleles for the Larsen score. Mann-whitney U tests comparing genotypes for the Larsen score. ORs to test proportion positive or negative RF & ACPA.

Gene/ rs number/ Severity marker/ Reference	Sample size	Evidence of association	No evidence of association	Analysis performed
		respectively, p=0.069) however there was no association with ACPA or RF.		
HLA-DRB1/ rs1410766, rs322812 and rs347117/ Joint morbidity/ (Min et al., 2010)	UNK	Found a high peak (LOD = 3.29; NPL Z = 4.07) near the HLA-DRB1 region on chromosome 6. The linkage at 6p24 at rs1410766 [LOD = 2.66; nonparametric linkage (NPL) Z = 3.23] was statistically significant. Two other regions also showed possible linkage peaks: chromosome 7q30 at rs322812 (LOD = 2.47; NPL Z = 3.39) and chromosome 15p34 at rs347117 (LOD = 1.95; NPL Z = 2.80).		Linkage association study
HLA-DRB1/ Shared Epitope/ Erosions, SJC, HAQ/ (Morgan et al., 2010)	1046 RA		No evidence HLA-DRB1 SE associated with prevalent erosions, SJC or HAQ. SE included number of copies (0, 1 or 2) of HLA-DRB1*0101, 0102, 0401, 0404, 0405, 0408, and 1001 and did not look individually at the SE genotypes	Logistic (presence/absence of erosions) and linear regression (SJC, HAQ) were used to undertake trend tests of association, which were adjusted for symptom duration
HLA-DRB1/ UNK/ Larsen, DAS28, ESR, SJC, TJC/ (Innala et al., 2008)	210 RA		HLA-DRB1 alleles were not related to radiological progression or inflammatory activity over time. Unknown grouping of HLA-DRB1 alleles	UNK
HLA-DRB1/ Shared epitope/ Sharp-van der Heijde method/ (Huizinga et al., 2005)	408 RA	Large differences were observed between ACPA positive and ACPA negative patients. No apparent association was observed between SE positivity and progression of joint damage in ACPA negative patients. In contrast, radiographic severity scores were higher among ACPA positive patients who were SE positive than among those		Plots of Sharp score over time split by ACPA positivity and SE status

Gene/ rs number/ Severity marker/ Reference	Sample size	Evidence of association	No evidence of association	Analysis performed
		who were SE negative.		
HLA-DRB1/ Shared epitope/ Sharp score/ (Suzuki et al., 2013)	830 RA	11.6% (95% CI: 4.1%-18.5%, p=0.0021, N=830) increase in joint damage, as measured by the Sharp score five years after disease diagnosis, for each copy of the RAA shared epitope motif that a patient has.		UNK
HLA-DRB1/ Shared epitope/ Sharp score/ (Mackie et al., 2012a)	3657 RA	Particularly in ACPA / RF positive disease, there was a hierarchy of severity associated with the SE alleles. The worst severity was associated with DRB1 *0404 or *0401 (p=0.0003 when compared to *0101 and *1001). For ACPA/RF positive only, there was a gene dose effect observed with a protective effect for D70 (OR 0.82, 95% CI 0.73–0.92, P=5.8 × 10−4). HLA-DRB1 SE alleles were also associated with ACPA-negative, RF-positive RA (OR 1.42 (1.15–1.76), P=0.0012).		UNK
HLA-DRB1/ Shared epitope/ Sharp score/ (Meyer et al., 2011)	143 RA	Shared epitope patients showed strong association with ACPA positive disease (OR = 10.2 and P = 0.0010, OR = 9.2 and P = 0.0028, respectively). Clinical scores and concentrations of the other biomarkers of disease activity were also generally higher in the shared epitope group vs no shared epitope group		UNK
HMOX1/ $(GT)_n$ repeat/ ESR, CRP, RF, DAS28, Ratingen radiographic damage scoring system/ (Wagener et al., 2008)	325 RA	After 9 years of follow up, subjects with short $(GT)_n$ repeat (n<25; SS genotype) had a better radiographic outcome that those carrying long (GT)n repeat (n>=25; LL genotype). Increase in damage score from SS to SL and LL is 8.2 (SE=6.1) and 12.1 (SE=5.9) respectively (p=0.047). Increase in damage score for allele L compared to S is 5.9 (SE=2.7), p=0.013).		Difference from baseline analysed using linear regression with dummy variables for genotype groups

Gene/ rs number/ Severity marker/ Reference	Sample size	Evidence of association	No evidence of association	Analysis performed
		Genotype LL. No effect seen for ESR, CRP, RF or DAS28.		
IL-1α/ rs17561 (+4845)/ Larsen wrist x-ray index/ (Jouvenne et al., 1999)	98 RA/ poly-arthritis	The percentage of patients carrying the rare IL-1A2 allele in the control population was 45%. This increased for destructive RA to 54.4% and decreased in non-destructive RA to 26.8%, (Destructive versus non-destructive, $p < 0.007$). All indices of disease activity and joint destruction were significantly lower in the patients positive for IL-1A1, and higher in those positive for IL-1A2		UNK
IL-1α/ rs1800587/ Presence of nodules, requirement for joint replacement and radiographic progression by Rau-Ratingen method/ (Harrison et al., 2008)	756 RA		No direct association between IL-1A (-889 C/A, rs1800587) and clinical severity characteristics	Modelled Logit transformation of radiographic score (ln[score/max score-score]) = square root of disease duration. Fishers exact test to compare allele frequency vs. mild (score<20) or severe (score>80)
IL-1α/ rs6712572, rs3783550, rs17561, rs378351, rs1800587, rs1894399, rs6746923, rs17597976/ With and without hand erosions/ (Johnsen et al., 2008)	712 RA patients test and 414 rep-licate		No robust, reproducible association between IL-1α variants and the proportion of patients with or without hand erosions.	Chi square test of the proportion with and without erosions. Analysed erosions at one time point and could be very dependent on aggressiveness of treatment received and length of treatment.
IL-1β/ rs1143634/ Presence or absence of erosive damage/ (Cantagrel et al., 1999)	108 RA	Carriage of the rare IL-1β +3954 (rs1143634) allele 2 was found to expose patients to an increased risk of erosive disease, with an OR of 8.20 (95% CI 2.59-25.84, $P < 0.0001$)	IL-1β -511 was not associated with the development of erosions	UNK

Gene/ rs number/ Severity marker/ Reference	Sample size	Evidence of association	No evidence of association	Analysis performed
IL-1β/ rs1143634/ Larsen wrist x-ray index/ (Buchs et al., 2001)	378 RA	Carriage of the rare IL-1β +3954 (rs1143634) allele 2 was associated with an increase in destructive arthritis as compared to non-destructive arthritis (OR 1.7, 95% CI 1.1-2.8, 49.0% vs. 35.9%). Patients carrying this allele had more destruction (Larsen wrist radiological index: mean +/- s.e.m., 2.1 +/- 0.2 vs. 1.6 +/- 0.1, P = 0.005; Steinbrocker functional index: 2.4 +/- 0.1 vs. 1.9 +/- 0.1, P = 0.002) and more active disease (Ritchie articular index: 8.1 +/- 0.8 vs. 5.3 +/- 0.6, P = 0.002; ESR: 36.6 +/- 2.9 mm/h vs. 25.3 +/- 1.8 mm/h, P = 0.002). This contribution was independent from that of HLA DR4/DR1 to severity		UNK
IL-1β/ rs1143634/ DAS28, ESR, SJC, TJC/ (Pawlik et al., 2005a)	93 RA	For patients carrying the rare IL-1β +3954 (rs1143634) allele 2, the active form of RA was more frequently diagnosed. Moreover in these patients the measurements of disease activity (DAS28 score, ESR, number of swollen and tender joints) were significantly increased		UNK
IL-1β/ rs1143634/ Cumulative disease score/ (Cvetkovic et al., 2002)	54 RA	Patients with genotype A2A2 of IL-1β had higher accumulated disease activity score than patients with A1A1 and A1A2 ($p < 0.05$)		UNK
IL-1β/ rs16944/ Larsen score/ (Genevay et al., 2002)	233 RA	IL-1β -511 (rs16944) allele 2 was associated with milder radiographic progression. The slope of Larsen progression in the rare allele groups diverged significantly from those of the frequent allele groups after approximately 20 years of disease duration ($P < 0.001$)		UNK
IL-1β / rs16944, rs1143623, rs4848306/	756 RA		No direct association between IL-1B (-511 A/G,	Modelled Logit transformation of radiographic score (ln[score/max

Gene/ rs number/ Severity marker/ Reference	Sample size	Evidence of association	No evidence of association	Analysis performed
Presence of nodules, requirement for joint replacement and radiographic progression by Rau-Ratingen method/ (Harrison et al., 2008)			rs16944), IL-1B (-1464 C/G, rs1143623), IL-1B (-3737 G/A, rs4848306) and clinical severity	score-score]) = square root of disease duration. Fishers exact test to compare allele frequency vs. mild (score<20) or severe (score>80)
IL-1β/ rs4849125, rs7596684, rs1143634, rs1143633, RA1, rs1143627, rs16944, RA3, rs13013349, rs13032029, RA4, rs4447608, rs6735739, rs6745746, rs12053091/ With and without hand erosions/ (Johnsen et al., 2008)	712 RA patients test and 414 rep-licate		No robust, reproducible association between IL-1β variants and the proportion of patients with or without hand erosions.	Chi square test of the proportion with and without erosions. Analysed erosions at one time point and could be very dependent on aggressiveness of treatment received and length of treatment.
IL-1β/ rs16944, rs1143634 / HAQ/ (Lubbe et al., 2008)	141 RA	IL-1β -511 and +3954 were found to be possible polymorphisms associated with disease severity in RA. Carriage of one copy of IL-1β-511 T allele was associated with worse MHAQ scores (corrected for disease duration) compared to patients not carrying this allele (mean=1.54 vs. 1.0, respectively, p=0.02)		Linear regression was used to quantify the relationship between severity markers and genotypes.
IL-1RN, IL-1RA/ IV allele/ Inflammatory activity/ (Huang et al., 2001)	104 RA	The IL-1RN IV allele was more common in patients with low inflammatory activity. In contrast, the IV allele of IL-1Ra was significantly increased in RA patients with low inflammatory activity (P=0.03)		UNK
IL-1RN/ Variable number of tandem repeats (VNTR)/ Presence or absence of erosive damage/	108 RA		The IL-RN variable number of tandem repeat was not associated with development of erosions	UNK

Gene/ rs number/ Severity marker/ Reference	Sample size	Evidence of association	No evidence of association	Analysis performed
(Cantagrel et al., 1999)				
IL-1RN/ +2018 C allele and VNTR-T long 3-6 repeats / HAQ/ (Lubbe et al., 2008)	141 RA	IL-1RN*2 (+2018 C allele and VNTR-T long 3-6 repeats) was found to be a marker of erosive joint damage in Black South Africans with RA		Linear regression was used to quantify the relationship between severity markers and genotypes.
IL-2/ -330/ DAS28, ESR, SJC, TJC/ (Pawlik et al., 2005a)	93 RA	In patients with the GG genotype of IL-2, the active form of RA was more frequently diagnosed and measurements of disease activity (DAS28 score, ESR, number of swollen and tender joints) were also significantly increased.		UNK
IL-4/ VNTR/ Larsen score/ (Genevay et al., 2002)	233 RA	The rare allele of the IL-4 VNTR was associated with less severe course, the rare allele groups diverged significantly from those of the frequent allele groups after approximately 20 years of disease duration ($P < 0.001$).		Regression analysis by allele group
IL-4/ rs2243250/ DAS28, ESR, SJC, TJC/ (Pawlik et al., 2005d)	94 RA	Parameters of disease activity (DAS28 score, ESR, number of swollen and tender joints) were significantly increased for patients carrying the IL-4 -590 T allele (rs2243250) (genotypes CT and TT) compared to the homozygous CC genotype		UNK
IL-4/ rs2243250 and VNTR/ Presence or absence of erosive damage/ (Cantagrel et al., 1999)	108 RA		The IL-4 promoter variant -590 (rs2243250) and the 70 bp VNTR were not associated with erosive RA	OR & 95% CI
IL-4/ rs2227284, rs2243263, rs2243267 / Larsen score/ (Marinou et al., 2008)	965 RA		IL-4 (rs2227284, rs2243263 and rs2243267) were not found to be associated with Larsen scores	Modified Larsen score differences were tested for associations with each candidate gene polymorphism using the nonparametric Kruskal-Wallis

Gene/ rs number/ Severity marker/ Reference	Sample size	Evidence of association	No evidence of association	Analysis performed
				test
IL-4R/ rs1801275 and rs1805010/ Larsen score/ (Marinou et al., 2008)	965 RA		IL-4R (rs1801275 and rs1805010) were not found to be associated with Larsen scores	Modified Larsen score differences were tested for associations with each candidate gene polymorphism using the nonparametric Kruskal-Wallis test
IL-4R/ rs1805010/ Erosive vs non erosive disease/ (Prots et al., 2006)	471 RA	There was a significant difference in the distribution of the IL-4R I50V (rs1805010) between patients with erosive and non-erosive disease (chi-square = 15.68, P = 0.0004). Bone erosions at 2 years after disease onset were present in 68.1% of patients homozygous for the V50 allele compared with 37.0% of patients homozygous for the I50 allele (OR 3.86, P < 0.0001)		ORs & 95% CI
IL-6/ rs1800795/ DAS28, ESR, SJC, TJC/ (Pawlik et al., 2005c)	98 RA	Carriers of the IL-6 -174 G (rs1800795) alleles had a significant increase in DAS28, ESR, SJC and TJC		UNK
IL-6/ rs1800795/ Joint radiographs within first 2 years after onset/ (Oen et al., 2005)	181 RA		Although the IL-6 genotype -174G/G was positively correlated with pain [regression coefficient B = 0.899, 95% confidence intervals (CI) 0.185, 1.612, P = 0.014] the IL-6 -174 had no significant effect in radiographic damage	UNK
IL-6/ rs1800795/ Larsen score/	964 RA	The IL-6 -174 G allele was associated with increasing radiographic damage (p=0.005) but a subgroup analysis showed this was only in		Modified Larsen score differences were tested for associations with each candidate

Gene/ rs number/ Severity marker/ Reference	Sample size	Evidence of association	No evidence of association	Analysis performed
(Marinou et al., 2007)		patients who were either RF positive (p=0.004) or ACPA positive (0.01)		gene polymorphism using the nonparametric Kruskal-Wallis test
IL-6/ rs1800795/ FMD%/ (Palomino-Morales et al., 2009)	311 RA	Homozygous IL-6 -174 GG genotype had more severe endothelial dysfunction (FMD%) than GC or CC genotypes. 4.2 vs. 6.3 vs. 6.0 respectively, p=0.02		UNK
IL-6R/ rs8192284/ DAS28/ (Lamas et al., 2010)	281 RA	A statistically significant interaction was observed between IL-6R rs8192284 polymorphism and the presence of ACPA (p=0.008). An inverse relationship between the polymorphism and DAS28 was observed depending on ACPA status		A mixed-effect model was used to analyse the measurements.
IL-6R/ rs8192284/ DAS28/ (Rodriguez-Rodriguez et al., 2011)	281 RA	DAS28 and plasma sIL-6R levels are positively correlated with ACPA positive patients (r=0.45, p=0.0336) and negatively correlated with ACPA negative patients (r=-0.45, p=0.0825).		UNK
IL-6R/ rs8192284/ Erosion score/ (Ceccarelli et al., 2011)	77	Evidence of association between IL-6 -174 and disease severity p=0.007.		UNK
IL-6R/ rs8192284/ Radiographic erosions/ (Gottenberg et al., 2012)	578 RA	Serum levels of IL-6 were found higher in subjects with radiographic progression at 1 year (OR 2.4, 95% CI: 1.1 to 5.2, p=0.005)		UNK
IL-10/ rs1800896/ Modified Sharp score/ (Huizinga et al., 2000)	138 RA	Patients with the IL-10 -1082 (rs1800896) GG genotype had increased radiographic damage score. Increase during the first 6 years was 9 +/- 9 per year for -1082AA genotype vs. 19 +/- 16 per year for patients with the -1082GG genotype		Regression of the mean increase in Sharp score

Gene/ rs number/ Severity marker/ Reference	Sample size	Evidence of association	No evidence of association	Analysis performed
		(P= < 0.02)		
IL-10/ -2849/ Sharp-van der Heijde method/ (Lard et al., 2003)	283 RA	Patients with the -2849 AG/GG genotype, which is associated with high IL-10 production, had higher autoantibody titres at baseline		UNK
IL-10/ rs1800896/ Presence or absence of erosive damage/ (Cantagrel et al., 1999)	108 RA		The IL-10 -1082 was not associated with the presence of erosive damage	OR & 95% CI
IL-10/ rs1800896, rs18000872 and -819 (rs UNK)/ Joint radiographs within first 2 years after onset/ (Oen et al., 2005)	181 RA		The IL-10 promoter polymorphisms -1082, -819 and -512 were not associated with radiographic damage	UNK
IL-10/ rs1800896 and rs18000872/ SJC, TJC, ESR, CRP, duration of morning stiffness/ (Pawlik et al., 2005b)	95 RA		The IL-10 variants -1082 and -592 are not genetic risk factors for RA severity	Correlation
IL-10/ rs18000872/ Larsen score/ (Marinou et al., 2007)	964 RA	Patients with the IL-10 −592CC genotype had more extensive radiographic damage than did those with the AC or AA genotype (*P* = 0.006), but this was observed only among patients who were RF negative (*P* = 0.002) or ACPA negative (*P* = 0.002)		Modified Larsen score differences were tested for associations with each candidate gene polymorphism using the nonparametric Kruskal-Wallis test, or Cuzicks test for trend if a trend was observed.
IL-10/ rs18000872/ DAS28, bone erosions, deformities, extra articular	336 RA	The A allele of IL-10 -592 polymorphism was found to be marginally associated with the higher DAS28 score (C allele, 5.57 +/- SE 1.19 vs. A allele 5.77 +/- SE 1.19, p=0.045). Other alleles		Multiple linear regression models for DAS28, logistic regression models for joint erosions, deformities or presence of extra

Gene/ rs number/ Severity marker/ Reference	Sample size	Evidence of association	No evidence of association	Analysis performed
features/ (Gambhir et al., 2010)		were not found to be significant and IL-10 -592 and -1082 did not associate with bone erosions, deformities or presence of extra articular features		articular features.
IL-15/ rs7667746, rs7665842, rs2322182, rs6821171 and rs4371699**/** Sharp score/ (Knevel et al., 2012b)	1318 RA	In a meta-analysis of 4 cohorts, subjects with the most frequent homozygous genotype for rs7667746, rs7665842, rs2322182, rs6821171 and rs4371699 had 0.94 ($p=4.0\times10^{-6}$), 1.04 ($p=3.8\times10^{-4}$), 1.09 ($p=5.0\times10^{-3}$), 1.09 ($p=5.0\times10^{-3}$) and 1.09 (p=9.4×10−3) fold rate of joint destruction compared to other patients respectively.		Meta-analysis
IL-17F/ rs763780, rs2397084/ TJC, SJC, HAQ, Labs, DAS28, Disease duration, CRP, VAS, RF, early-late RA, ACPA, gender/ (Paradowska-Gorycka et al., 2010)	220 RA	IL-17F gene His161ARG 7488 A/G rs763780 variant had some evidence of association with TJC, creatinine, HAQ and DAS28 but not with SJC, disease duration, CRP, VAS, platelets, haemoglobin, RF, gender, early/late RA and ACPA Some evidence of IL-17F gene Glu126Gly 7383 A/G rs2397084 may be associated with longer disease duration but not statistically significant and no association with other disease activity measurements.		Wilcoxon test and chi-square test with yate's correction
MBL/ 0/0 genotype/ Larsen score, 30% of maximum radiographic destruction (RE30), ESR, SJC, Labs/ (Graudal et al., 2000)	140 RA	Patients with the MBL defective 0/0 genotype, which is associated with undetectable levels of plasma MBL, had worst radiographic outcome. RE30 was 3.1 (95% CI 1.8-5.1) in the MBL-insufficient group versus the MBL-competent group ($P < 0.0001$). RE30 occurred in 50% of MBL-competent patients within 17 years, while such an event occurred 9 years earlier in MBL-insufficient		MBL-insufficient patients (those with 2 defective structural MBL alleles or with 1 defective allele combined with a low-expression variant of the normal allele). Relative risks of defection vs. not defective against severe radiographic outcome which was defined as 30% of maximum

Gene/ rs number/ Severity marker/ Reference	Sample size	Evidence of association	No evidence of association	Analysis performed
		patients (i.e., within 8 years) (P < 0.0001). During the first 15 years, there was a significant trend toward lower haemoglobin levels (P < 0.04), higher ESRs (P< 0.02), and a higher number of swollen joints (P < 0.05) in the MBL-insufficient group.		radiographic destruction (RE30)
MBL/ 0/0 genotype/ Presence or absence of erosions/ (Ip et al., 2000)	211 RA	Patients with erosive and serious extra-articular disease had significantly lower serum MBL levels than those without. Significantly more patients with erosive disease had a codon-54 mutation of the MBL gene compared with those with non-erosive disease. Serum MBL levels did not correlate with drug treatment or with disease activity.		UNK
MBL/ -550/ DAS score/ (Gupta et al., 2005)	120 RA	The promoter polymorphism at -550 of the minor allele G was observed significantly more frequently in severe RA patients compared with the less severe group (P=0.003). The haplotype LYA was significantly more frequent in the less severe group (P=0.03) and haplotype HYA was significantly more frequent in the severe RA patients (P=0.04).		UNK
MBL/ UNK/ Larsen score/ (Jacobsen et al., 2001)	68 RA	Patients with early polyarthritis homozygous for MBL variant alleles had an increased risk of having erosive RA at inclusion by a factor of 4.7 (p = 0.02) and after one year by a factor of 3.6 (p = 0.04). MBL deficiency was associated with increased levels of CRP and IgM RF at inclusion (p < 0.05).		UNK
MBL/ -221 (XX, XY, YY) MBL2/ ACPA, DAS28, HAQ, Sharp score/	158 RA	High scores of disease activity, CRP-based DAS28 (p=0.02), and physical disability by HAQ (p=0.01) were associated with high MBL2 expression		UNK

Gene/ rs number/ Severity marker/ Reference	Sample size	Evidence of association	No evidence of association	Analysis performed
(Jacobsen et al., 2009)		genotypes in a gene-dose dependent way, but only in ACPA positive patients. At this early stage of the disease there was no association with erosion score from radiographs.		
MBL/ -550, -221, codon 52 and codon 54/ Larsen score/ (Barton et al., 2004)	438 inflam-matory polyar-thritis		None of the SNPs (positions -550, -221, codon 52 and codon 54) was associated with development of erosions or Larsen score at 1 year and 5 years	Allele frequencies by patients with or without erosions in addition to an analysis of change in Larsen score (method of analysis UNK).
MBL/ 550, -221, codon 52 and codon 54/ Need for anti-TNF therapy / (van de Geijn et al., 2008)	639 RA		No association between MBL groups and disease severity	UNK
Mediterranean FeVer (MEFV) gene mutations/ Five frequent mutations (E148Q, M694V, M694I, M680I, V726A) and three rare mutations (A744S, R761H and P369S) DAS28, CRP, Larsen score/ (Koca et al., 2010)	103 RA	Deformed joint count was significantly higher in the mutation carrier group (mean number of deformed joints =6.2, SE=9.7) than those of the non-carrier group (mean number of deformed joints =2.6, SE=5.1) in chr 16p13.3 (p=0.026). The level of CRP, DAS28 and modified-Larsen scores were slightly but not significantly higher in the mutation carrier group in chr 16p13.3		T-tests to compare groups, fisher's exact to compare categorical variables, and ORs (95%CI) for the assessment of risk factors. Analysis of covariance was used to adjust variables for disease durations.
Metalloproteinase (MMP)-1 and MMP-3/ 1G/2G(MMP1) and 5A/6A (MMP3) Ratingen score/ (Dorr et al., 2004)	308 RA	There was a significant effect on the degree of radiographic joint destruction with the 1G-5A haplotype (P = 0.0001) and the interaction term 'estimated number of 1G-5A haplotypes x duration of disease' (P = 0.0007). The interaction revealed that the 1G-5A haplotype had a protective effect over a period of about 15 years of RA, but was associated with a worse		Factorial regression containing interaction terms & main effects (allele frequency * disease duration).

Gene/ rs number/ Severity marker/ Reference	Sample size	Evidence of association	No evidence of association	Analysis performed
		radiographic progression in later years. Similar results were also found with the 1G allele of MMP1 alone (P = 0.015), interaction '1G x duration of disease (P = 0.014).		
MMP-3/ 5A/6A biallelic polymorphism/ Sharp-van der Heijde method/ (Constantin et al., 2002)	103 RA	The MMP-3 6A/6A genotype was associated with the highest Sharp score both at baseline and after a 4 year follow-up and with the highest progression of the Sharp score over the 4 years of follow-up. Patients homozygous for MMP-3 6A and DRB1 SE had the highest progression of the Sharp score.		UNK
MMP-3/ 6A polymorphism/ Larsen score/ (Mattey et al., 2004)	254 RA	Patients homozygous for the MMP-3 6A allele had more radiographic damage (measured by Larsen score) than those with other genotypes (109.8 vs. 91.1, P=0.04). Patients with the 6A/6A genotype also had more functional impairment and higher serum pro MMP-3 levels, although only the latter was significant (P=0.002).		UNK
MTHFD1 / rs17850560/ DAS28/ (Grabar et al., 2010)	154 RA patients	After exclusion of patients co-treated with other DMARDs (not methotrexate), a significant association of MTHFD1 1958GG genotype was found with lower disease activity (p=0.021, OR=4.674, 95% CI 1.266 to 17.262)		Disease activity low (DAS28 <=3.2) vs. moderate/high (DAS28 >3.2). ORs, 95% CI and p-value.
P53/ Coden 72/ Sharp-van der Heijde method/ (Macchioni et al., 2007)	122 RA	At five years, patients carrying the Pro/Pro genotype compared to the Arg/Arg genotype, had a significantly higher percentage of eroded joints (Pro/Pro 93%, Arg/Arg 52%, p=0.0001), mean number of eroded joints per patient (Pro/Pro 13.2, Arg/Arg 3.6, p=0.0001). The mean Sharp score, joint space narrowing score and total damage score were significantly higher in the Pro/Pro subgroup compared with the		UNK

Gene/ rs number/ Severity marker/ Reference	Sample size	Evidence of association	No evidence of association	Analysis performed
		Arg/Arg and Arg/Pro subgroups.		
Peptidyl arginine deiminase type 4 (PADI4)/ rs2240340/ DAS28, cumulative therapy intensity, steinbrocker score/ (Hoppe et al., 2009)	373 RA	PADI4 genotype, C>T rs2240340 was associated with ACPA status, SE, Anti-nuclear antibodies and disease activity. T allele exhibited a significant trend towards higher Steinbrocker scores (<=I vs. II vs. III vs. IV) testing the frequency of T allele across the groups p<0.004 and for the T/T genotype p=0.008 when adjusted for covariates		Non-erosive (steinbrocker score <=1 and erosive (steinbrocker score II-IV). ORs & 95% CI by logistic regression adjusting for covariates. Cuzick non-parametric test for trend.
PADI4/ Anti-hPAD4/ DAS28, Sharp-van der Heijde method/ (Halvorsen et al., 2009)	40 RA	Anti-hPAD4 positive patients had more severe disease (DAS28 and Sharp score) than the negative patients at baseline (p=0.049 and p=0.047 respectively) and after 1 year (on anti-TNF-alpha therapy) (p=0.016 and p=0.032). The mean change in erosion score from baseline to 1 year were 1.27 (0.11 to 2.44) vs. -0.32 (-1.17 to 0.54) (p=0.023).		Anti-hPAD4 levels compared using Wilcoxon rank sum. Change in Sharp score assessed using Wilcoxon signed rank tests. Erosive progression tested using fishers exact and bivariate exact logistic regression
PADI4/ padi4_89 (rs11203366) padi4_90 (rs11203367), padi4_92 (rs874881) / Sharp-van der Heijde/ (Harris et al., 2008)	129 RA	Anti–PAD-4 negative patients (n=83) with those with anti–PAD-4 autoantibodies scored as 3+ in the immunoprecipitation assay (n =26). Mean unadjusted Sharp scores were 57 (95% CI 43.6–70.9) in the anti–PAD-4 negative group, compared with 132 (95% CI 90.6–173.7) in the group with high anti–PAD-4 scores. These differences were statistically significant (*P* <0.001) and remained significance (P=0.001) after adjusting for confounding effects.		ANOVA of mean Sharp scores with and without potential confounding variables in the model
PADI4/ rs2240340/ Japanese HAQ score, disease durations/ (Nishimoto et al., 2008)	1384 RA		No evidence that PADi-94 is a predictor of aggravation of functional impairment of RA in the Japanese population	Linear regression analysis on the HAQ scores and ANOVA on the disease duration.
PADI4/	830 RA	7.3% (95% CI: 0.14%-15%, p=0.037, N=830)		Meta-analysis

Gene/ rs number/ Severity marker/ Reference	Sample size	Evidence of association	No evidence of association	Analysis performed
rs2240340/ Sharp score/ (Suzuki et al., 2013)		increase in joint damage, as measured by the Sharp score five years after disease diagnosis, for each copy of the rare T allele		
PTPN22/ rs2476601/ Sharp-van der Heijde method/ (Lie et al., 2007)	238 RA	An association between annual progression rate of Sharp-van der Heijde score and PTPN22 R620W (rs2476601) T-allele carriers was found, ($p = 0.01$), which was also present when only patients positive for the shared epitope were analysed ($p = 0.03$).		UNK
PTPN22/ rs2476601/ Larsen score/ (Marinou et al., 2007)	964 RA	Marginally significant +1858 T allele association with radiological damage (Median Larsen score TT=50, TC=33, CC=25, Cuzicks test for trend $p=0.04$)		Cuzicks test for trend
PTPN22/ rs2476601/ Larsen score/ (Pierer et al., 2006)	123 RA		No significant differences in disease activity or Larsen scores were detected	Mann-Whitney U test or t-test.
PTPN22/ rs2476601/ Erosive damage/ (Steer et al., 2005)	302 RA		No evidence of an association between PTPN22 and the presence or absence of erosive damage	Presence or absence of erosive disease testing CC/TT genotypes by ORs, 95% CIs.
PTPN22/ rs2476601/ Larsen, DAS28, ESR, SJC, TJC/ (Innala et al., 2008)	210 RA		PTPN22 T variant alleles were not related to radiological progression or inflammatory activity over time	UNK
PTPN22/ rs2476601/ Erosions, SJC, HAQ/ (Morgan et al., 2010)	1046 RA		No evidence PTPN22 is associated with prevalent erosions, SJC or HAQ.	Logistic (presence/absence of erosions) and linear regression (SJC, HAQ) were used to undertake trend tests of association, which were adjusted for symptom duration
PTPN22/	2680 -	Meta-analysis of six cohorts (N=2680) found	No evidence of an increase or	Fixed Mantel Haenszel and

Gene/ rs number/ Severity marker/ Reference	Sample size	Evidence of association	No evidence of association	Analysis performed
rs2476601/ ACPA status and Larsen score/ (Taylor et al., 2013)	3172 RA	smoking and the PTPN22 genotype increased the risk of ACPA positive disease particularly in combination with each other (OR=2.22, 95%CI 1.69-2.91, p=8.3 x10-9).	decrease in risk of erosions despite association between ACPA positive disease and erosive damage	random effects Dersimmion & Laird Meta analyses.
Stromal cell–derived factor 1 (SDF-1 or CXCLI2)/ 3'-UTR (801 G/A)/ Sharp-van der Heijde/ (Joven et al., 2005)	138 RA	10 years after RA disease diagnosis patients with the SDF-1 3'-UTR AA genotype had an increase in mean Sharp/van der Heijde score of 13.7 compared to 8.2 for patients with the GG or GA genotypes considered together (*P* = 0.020)		ANOVA
T cell receptor (TCR)/ 2kb allele/ Modified Sharp method/ (Devries et al., 1993)	111 RA	Radiographic progression (modified Sharp method) after a three year follow up, was significantly less in the group possessing the 2.0 kb allele (p=0.03).		UNK
TGFβ/ +869/ Larsen score/ (Mattey et al., 2005)	208 RA	Patients carrying a TGFβ1 +869 T allele had a higher mean HAQ score than those without this allele (1.60 *v* 1.22, p = 0.04). The T allele was also associated with higher five year mean area under the curve (MAUC) ESR, and nodular disease. Larsen score was higher in patients with the TT genotype compared with CC + CT genotypes, although this was not significant after correction for disease duration. There was a trend of increasing mortality risk with T allele dose after adjustment for age, sex, and disease duration. Hazard ratio = 1.6 (95% CI, 1.1 to 2.4), p = 0.01.		The association of genotypes with normally distributed outcome measures (Larsen score) was assessed using Analysis of covariance with disease duration as a covariate. Association between genotypes and non-parametric data such as HAQ, MAUC, ESR, and C reactive protein was assessed using Kruskal–Wallis one way ANOVA on ranks. Cox-regression was used for survival analysis.
TGFβ/ -509/ Modified Sharp score/ (Kim et al., 2004)	143 RA	The progression of radiographic severity, which was defined by a modified Sharp score plotted against disease duration, was significantly faster in the carrier of T allele at the -509 (p=0.048).		Plotted Sharp score against disease duration
TGFβ/ +869/	77 RA	For ACPA positive patients, the TGFβ+869 TT genotype was associated with a lower total		UNK

Gene/ rs number/ Severity marker/ Reference	Sample size	Evidence of association	No evidence of association	Analysis performed
Erosions score/ (Ceccarelli et al., 2011)		erosion score (p=0.011). However in ACPA negative patients, the TGFβ+869 TT genotype showed a trend towards a higher total erosion score (p>0.05).		
TGFβ/ codon 10T →C and codon 25G →C / Joint radiographs within 2 years of onset/ (Oen et al., 2005)	181 RA	The homozygous TGF-β1 codon 25G/G genotype showed a protective effect against joint space narrowing on radiographs taken within 2 years of disease onset, but confidence intervals were wide [OR 0.176, 95% CI: 0.037 to 0.837 *P* = 0.029]. The TGFβ1 10T→C variant shows no association.		Kruskal–Wallis, χ^2 or Fisher's exact tests
TNF-alpha/ rs1800629 -308(G/A)/ Cumulative disease score/ (Cvetkovic et al., 2002)	54 RA	Patients having the genotype A1A2 of TNF-alpha developed more severe disease compared with patients with A1A1 genotype: they were younger at disease onset ($p < 0.05$), had a higher accumulated disease activity ($p < 0.05$) and worse functional class ($p < 0.05$)		UNK
TNF-alpha/ rs1800629 -308/ Steinbrocker radiographic score, HAQ, standard disability index / (Nemec et al., 2008)	130 RA	Significant differences observed in radiographic progression of disease based on the Steinbrocker radiographic score (p=0.03) and functional ability (HAQ) (p=0.03) suggesting an association of the -308 G/A polymorphism of the TNF-alpha gene with the severity of RA.		Erosive (steinbrocker score stadium II-IV) vs. non-erosive (steinbrocker score stadium I) and standard disability index (<=1 vs. >1) compared using ORs, 95% CI's and fishers exact test
TNF-alpha/ rs1800629 -308/ DAS28/ (Gambhir et al., 2010)	222 RA		-308(G/A) and -863(C/A) of TNF gene did not associate with DAS28, bone erosions, deformities or presence of extra articular features	Multiple linear regression models. Logistic regression models for bone erosions, deformities or presence of extra articular features.
TNFAIP3/ OLIG3 rs6920220, rs10499194, rs675520, rs9376293, rs1878658/	324 RA	rs6920220 (A) and rs10499194 (C) lie close to TNFAIP3 and were found to be associated with ACPA positive disease in long standing RA but this was not replicated in a study with shorter		Average increase in sharp-van der Heijde scores during the follow-up period was estimated for each person by regression

Gene/ rs number/ Severity marker/ Reference	Sample size	Evidence of association	No evidence of association	Analysis performed
Sharp-van der Heijde scores, ACPA (Scherer et al., 2010)		duration of RA. rs675520 (G) found to be significantly associated with increase in Sharp score (median slope AG/GG=4.6, AA=2.3; Mann-Whitney p=0.007) rs9376293 (C) associated with increase in Sharp score (median slope CC/CT=4.5, TT=3.0; Mann-Whitney p=0.021) No significant influence of rs1878658, rs10499194 and rs6920220 were found on radiographic joint damage		analysis. Subsequently the average increase (slope) of scores for each genotype was compared non-parametrically using the Mann-Whitney rank-sum test.
TNFAIP3/ OLIG3 rs6920220/ Erosions, SJC, HAQ/ (Morgan et al., 2010)	1046 RA		No evidence OLIG3/TNFAIP3 associated with prevalent erosions, SJC or HAQ	Logistic (presence/absence of erosions) and linear regression (SJC, HAQ) were used to undertake trend tests of association, which were adjusted for symptom duration.
TNFAIP3/ rs2230926/ Sharp score/ (Suzuki et al., 2013)	830 RA		No evidence of association between rs2230926 and Sharp scores.	Meta analysis
TNF receptor/ -383 TNFRI/ DAS28/ (Valle et al., 2010)	190 RA	The TNFRI -383 A/A genotype carriers had higher DAS28 score than A/C genotype (p=0.02)		Means comparisons were evaluated using the Mann-Whitney U test.
TNFSF1b/ rs1061622/ 3 and 6 year follow up joint damage/ (Toonen et al., 2008)	248 RA	TNF receptor super family 1b (TNFSF1b) gene (rs1061622) (676T>G, M196R) shows a significant difference in progression of radiological joint damage between the 3 genotype groups (TT, TG and GG) after 3 years follow-up (p=0.02), which lost significance after adjustment for multiple testing (p=0.06) and was also not significant at 6 years of follow up		Linear regression modelling of the mean joint damage

Gene/ rs number/ Severity marker/ Reference	Sample size	Evidence of association	No evidence of association	Analysis performed
		(p=0.29).		
Toll-like Receptors (TLR)/ rs5741883/ RF positivity, DAS28, joint damage (Rau scores) at 3 and 6 months/ (Enevold et al., 2010)	319 RA	After Bonferroni correction, there was a moderate association between RF positivity and TLR8 (rs5741883).	None of the 22 SNPs in TLR2, 3, 4, 5, 7, 8, and 9 had a statistically significant association with any RA clinical characteristics	Analysed for association
Toll-like Receptors (TLR)/ rs4986790, rs4986791/ Larsen score/ (Sheedy et al., 2008)	965 RA		TLR4 299 (rs4986790) and TLR4 399 (rs4986791) do not contribute to RA severity	Kruskal-Wallis test of the median Larsen score
TNF receptor 1 (TRAF1)/ rs10818488/ Sharp-van der Heijde method/ (Kurreeman et al., 2007)	268 RA	Carriers of the minor susceptibility A allele of rs10818488 had a two- fold higher radiological damage at 2 years after inclusion (p=0.008).		UNK
TRAF1/ rs10760130/ Erosions, SJC, HAQ/ (Morgan et al., 2010)	1046 RA	Some evidence that TRAF1/C5 (rs10760130) A allele was associated with more severe HAQ scores when adjusted for symptom duration. Mean (95% CI) AA=1.25 (1.11-1.38), GA=1.34 (1.24-1.43), GG=1.50 (1.35-1.65), OR(95% CI) =0.09 (0.01-0.19) p=0.031. No evidence TRAF1/C5 associated with prevalent erosions or SJC		Logistic (presence/absence of erosions) and linear regression (SJC, HAQ) were used to undertake trend tests of association, which were adjusted for symptom duration.
Tristetraprolin (TTP)/ 359(A/G), -503(A/C)/ Using of Infliximab/ (Suzuki et al., 2008)	155 RA	TTP (also known as Tis11, Nup475 and GOS24) located on Chr 19, position -686 to +25, 2 SNPS at -359(A/G), -503(A/C) were investigated and found to mildly affect promoter activity (allele A had a 1.5-2 fold increase in Luciferase activity than that with allele G, p<0.001) and thus may influence the disease activity.		Fisher's exact test of the use of infliximab vs. genotype.
Other region/ rs322812 and rs347117/	UNK	Two other regions also showed possible linkage peaks with joint morbidity: chromosome 7q30 at		Linkage analysis study. The phenotypic variables analysed

Gene/ rs number/ Severity marker/ Reference	Sample size	Evidence of association	No evidence of association	Analysis performed
Joint morbidity, RF/ (Min et al., 2010)		rs322812 (LOD = 2.47; NPL Z = 3.39) and chromosome 15p34 at rs347117 (LOD = 1.95; NPL Z = 2.80).		were the level of RF and score on the Joint Alignment and Motion (JAM) scale. The scale was modified by dividing by RF values relevant to disease severity.
Other region/ Various/ DAS28/ (Junta et al., 2009)	23 RA	Disease activity modulated the expression of 106 genes of which 91 were exclusively observed in RA patients exhibiting active disease (DAS28>5.0). The functions of these genes were related to signal transduction, apoptosis, response to stress, immune response and response to DNA damage stimulus. The remaining 15 genes had their expression influenced by the presence of SE (HLA-DRB1 *0401, *0404, *0405, *0408, *0101, *0102, *1001 and *1402) and ACPA		Analysis of the data using the significance analysis of microarrays algorithm together with a Venn diagram allowed the identification of shared and of exclusively modulated genes, according to patient features.

Appendix C: Description of SNPs modelled in the 'all subjects' dataset

After exclusion of SNPs with >20% subjects with missing data, the remaining 368 SNPs were distributed throughout the genome as follows:

Chromosome	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23
Frequency of SNPS	31	19	9	26	19	92	14	6	15	15	11	20	9	3	4	9	11	9	4	3	3	34	2

Here is a list of the 368 SNPs used in modelling.

rs333, rs689, rs6311, rs6313, rs6314,rs7343, rs8873, rs16944, rs17561, rs26232,rs26510, rs30187, rs30245, rs42041, rs42046,rs84458, rs182429, rs195967, rs213950, rs220704,rs228934, rs228935, rs228937, rs228945, rs228947,rs228954, rs228975, rs228979, rs229528, rs229541,rs231707, rs231735, rs231775, rs380421, rs394581,rs443198, rs472391, rs473892, rs508214, rs540386,rs548234, rs550523, rs553247, rs584794, rs590523,rs597846, rs609438, rs617956, rs626787, rs632020,rs632535, rs633010, rs636393, rs653178, rs665668,rs667520, rs678385, rs706778, rs719149, rs719150,rs729749, rs730560, rs743776, rs743777, rs743778,rs743779, rs758664, rs763361, rs775241, rs775249,rs791590, rs805292, rs842647, rs854350, rs864745,rs873308, rs874040, rs892188, rs896135, rs917997,rs923658, rs929230, rs932744, rs951005, rs983230,rs1003693, rs1003694, rs1054028, rs1076933, rs1078109,rs1160542, rs1295686, rs1304037, rs1372884, rs1398553,rs1410160, rs1436272, rs1445898, rs1447888, rs1464510,rs1465788, rs1510702, rs1545092, rs1545783, rs1556837,rs1678542, rs1716157, rs1773560, rs1793004, rs1799724,rs1800629, rs1800925, rs1801275, rs1805010, rs1837519,rs1872779, rs1878658, rs1887346, rs1980422, rs1990760,rs2002842, rs2004640, rs2009345, rs2027276, rs2069311,rs2069762, rs2069777, rs2069778, rs2071592, rs2073839,rs2075800, rs2104286, rs2188776, rs2210918, rs2227284,rs2230926, rs2235330, rs2240340, rs2242653, rs2243263,rs2243267, rs2250259, rs2256965, rs2268146, rs2276418,rs2281094, rs2292239, rs2327832, rs2339898, rs2431697,rs2476601, rs2514189, rs2542151, rs2568127, rs2569702,rs2618476, rs2666236, rs2677821, rs2715038, rs2736340,rs2771369, rs2793108, rs2812378, rs2816316, rs2837960,rs2844479, rs2872507, rs2888334, rs2900180, rs2941794,rs3087243, rs3087456, rs3093023, rs3118469, rs3176767,rs3194051, rs3218253, rs3218258, rs3218292, rs3218312,rs3218315, rs3218316, rs3218322, rs3218339, rs3748816,rs3757173, rs3761959, rs3777612, rs3788013, rs3803815,rs3807306, rs3816587, rs3816769, rs3825932, rs4112788,rs4133002, rs4265819, rs4272626, rs4492018, rs4505848,rs4535211, rs4637409, rs4675600, rs4677742, rs4695391,rs4750316, rs4755453, rs4760169, rs4777183, rs4791034,rs4810485, rs4833248, rs4869816, rs4892117, rs4895499,rs4896286, rs4896303, rs4963128, rs4986790, rs4986791,rs5029394, rs5029937, rs5029938, rs5029939, rs5754217
,rs5756391, rs5980742, rs5995385, rs6000570, rs6017667,rs6441961, rs6473517, rs6490130, rs6568431, rs6570184,rs6682654, rs6822844, rs6897932, rs6903624, rs6909753,rs6918078, rs6920220, rs6927172, rs6932056, rs6933404,rs7021049, rs7026551, rs7041422, rs7091432, rs7234029,rs7257520, rs7313599, rs7543174, rs7574865, rs7579737,rs7601303, rs7722135, rs7749323, rs7752903, rs7753873,rs7766288, rs7949682, rs8045689, rs8084582, rs8177374,rs8180663, rs8192284, rs9270657, rs9321627, rs9359049,rs9366826, rs9389526, rs9389541, rs9402914, rs9402927,rs9494850, rs9550642, rs9564915, rs9565072, rs9622555,rs9770242, rs10015924, rs10036748, rs10040327, rs10282458,rs10489265, rs10499196, rs10499197, rs10516487, rs10760129,rs10760130, rs10785333, rs10786617, rs10812655, rs10814339,rs10818488, rs10865035, rs10878089, rs10905518, rs10910099,rs10914783, rs10917214, rs10919563, rs10954213, rs10997632,rs11080287, rs11151064, rs11162922, rs11203203, rs11203368,rs11586238, rs11651303, rs11718592, rs11724582, rs11732095,rs11755393, rs11755527, rs11761231, rs11762801, rs11868854,rs11876710, rs11970361, rs12035407, rs12094648, rs12137270,rs12150054, rs12194870, rs12194935, rs12198924, rs12206392,rs12251833, rs12403075, rs12525464, rs12527282, rs12527578,rs12551429, rs12603665, rs12708716, rs12722489, rs12723859,rs12746613, rs12840573, rs13017599, rs13031237, rs13061519,rs13119723, rs13192841, rs13207033, rs13242262, rs13277113,rs13315591, rs13393256, rs16881910, rs17005786, rs17015108,rs17037696, rs17066681, rs17085170, rs17143115, rs17223208,rs17388568, rs17534243, rs17612952, rs17696736, rs17780429,rs17810546, rs28665122, rs61330082

Appendix D: Amended 'spls' and 'valid' functions from mixOmics version 3.0 (amended version 4 for Larsen score and one component)

```
################################################################################
#    Here are two macros which were taken from the Mixomics Version 3.0 function##
#    and adapted numerous times for specific use on the SNP data modelling the  ##
#    Larsen score                                                               ##
#                                                                               ##
#  Version 1: The first updates were to output more of the datasets to be       ##
#     available for use after the macro was run                                 ##
#    For spls to spls_lyn, the changes enabled a list of the selected variables ##
#    ranked for each CV to be output into a dataset so once the model           ##
#    is selected in 8/10 folds in 50 runs, the variables can be                 ##
#    ranked in order of importance                                              ##
#                                                                               ##
#    For valid to splscv, The changes enable it to output the number of times   ##
#    each variable is selected in the M folds (Svar) and it outputs (Ycorr)     ##
#    Ycorr has the original Y values and the predicted Ys from the separate     ##
#    folds of the data                                                          ##
#                                                                               ##
# Version 2: Was written for the GWAS modelling                                 ##
#    The change was to make SNPs of 0 variance in some cross-validations just   ##
#    be excluded from that run to prevent the analysis failing                  ##
#                                                                               ##
# Version 3: amended from using 0 variance to include close to 0, as the NIPALS ##
#    algorithm could impute 0.0001 as SNP value but this was still insufficient ##
#    variation for the model to fit                                             ##
#                                                                               ##
# Version 4: amended code to export list of variable selection order instead    ##
#    of ranking top X variables and assigning equal last rank for those not     ##
#    selected                                                                   ##
#                                                                               ##
# To ensure consistency, the macros were moved to this program                  ##
# to be run at the start of any analyses program thus ensuring the same code    ##
# is being used each time                                                       ##
################################################################################

################################################################################
#  Original macros are from the following package: Mixomics Version 3.0         #
#                                                                               #
#  LÊ CAO, K.-A., I, G. & S, D. 2009. integrOmics: an R package to unravel      #
#  relationships between two omics data sets. Bioinformatics, 25 (21), 2855-2856.#
#  NOTE: the package 'integrOmics' has been renamed 'mixOmics'.                 #
#                                                                               #
# GONZÁLEZ, I., LÊ CAO, K.-A. & DÉJEAN, S. 2011. mixOmics: Omics Data Integration#
# Project. URL: http://www.math.univ-toulouse.fr/~biostat/mixOmics/.            #
################################################################################

################################################################################
#  MACRO 1: AMENDED spls FUNCTION and renamed spls_lyn                          #
################################################################################

spls_lyn <-
function(X,
         Y,
         ncomp = 2,
         mode = c("regression", "canonical"),
         max.iter = 500,
         tol = 1e-06,
         keepX = rep(ncol(X), ncomp),
         keepY = rep(ncol(Y), ncomp),
         ...)
{

    #-- validation des arguments --#
    if (length(dim(X)) != 2)
        stop("'X' must be a numeric matrix.")

    X = as.matrix(X)
    Y = as.matrix(Y)
```

```
  if (!is.numeric(X) || !is.numeric(Y))
      stop("'X' and/or 'Y' must be a numeric matrix.")

  n = nrow(X)
  q = ncol(Y)

  if ((n != nrow(Y)))
      stop("unequal number of rows in 'X' and 'Y'.")

  if (is.null(ncomp) || !is.numeric(ncomp) || ncomp <= 0)
      stop("invalid number of variates, 'ncomp'.")

  nzv = nearZeroVar(X, ...)

  if (length(nzv$Position > 0)) {
      warning("Zero- or near-zero variance predictors.
Reset predictors matrix to not near-zero variance predictors.
See $nzv for problematic predictors.")
      X = X[, -nzv$Position]
  }

     p = ncol(X)

  ncomp = round(ncomp)
  if(ncomp > p) {
      warning("Reset maximum number of variates 'ncomp' to ncol(X) = ", p, ".")
      ncomp = p
  }

  if (length(keepX) != ncomp)
      stop("length of 'keepX' must be equal to ", ncomp, ".")

  if (length(keepY) != ncomp)
      stop("length of 'keepY' must be equal to ", ncomp, ".")

  if (any(keepX > p))
      stop("each component of 'keepX' must be lower or equal than ", p, ".")

  if (any(keepY > q))
      stop("each component of 'keepY' must be lower or equal than ", q, ".")

  mode = match.arg(mode)

  #-- initialisation des matrices --#
  X.names = dimnames(X)[[2]]
  if (is.null(X.names)) X.names = paste("X", 1:p, sep = "")

  if (dim(Y)[2] == 1) Y.names = "Y"
  else {
      Y.names = dimnames(Y)[[2]]
      if (is.null(Y.names)) Y.names = paste("Y", 1:q, sep = "")
  }

  ind.names = dimnames(X)[[1]]
  if (is.null(ind.names)) {
      ind.names = dimnames(Y)[[1]]
      rownames(X) = ind.names
  }

  if (is.null(ind.names)) {
      ind.names = 1:n
      rownames(X) = rownames(Y) = ind.names
  }

  #-- centrer et réduire les données --#
  X = scale(X, center = TRUE, scale = TRUE)
  Y = scale(Y, center = TRUE, scale = TRUE)

  X.temp = X
  Y.temp = Y
  mat.t = matrix(nrow = n, ncol = ncomp)
```

```
    mat.u = matrix(nrow = n, ncol = ncomp)
    mat.a = matrix(nrow = p, ncol = ncomp)
    mat.b = matrix(nrow = q, ncol = ncomp)
    mat.c = matrix(nrow = p, ncol = ncomp)
    mat.d = matrix(nrow = q, ncol = ncomp)
       n.ones = rep(1, n)
       p.ones = rep(1, p)
       q.ones = rep(1, q)
       na.X = FALSE
    na.Y = FALSE
    is.na.X = is.na(X)
    is.na.Y = is.na(Y)
       if (any(is.na.X)) na.X = TRUE
    if (any(is.na.Y)) na.Y = TRUE
```

```
#  added by Lyn into version 1: created a blank matrix called ordselvar to complete the
ordering of selected variables by component #
   ordselvar=matrix(nrow=p, ncol=ncomp)

    #-- boucle sur h --#
    for (h in 1:ncomp) {
        nx = p - keepX[h]
        ny = q - keepY[h]

        #-- svd de M = t(X)*Y --#
        X.aux = X.temp
        if (na.X) X.aux[is.na.X] = 0

        Y.aux = Y.temp
        if (na.Y) Y.aux[is.na.Y] = 0

        M = crossprod(X.aux, Y.aux)
        svd.M = svd(M, nu = 1, nv = 1)
        a.old = svd.M$u
        b.old = svd.M$v

        #-- latent variables --#
        if (na.X) {
            t = X.aux %*% a.old
            A = drop(a.old) %o% n.ones
            A[is.na.X] = 0
            a.norm = crossprod(A)
            t = t / diag(a.norm)
            t = t / drop(sqrt(crossprod(t)))
        }
        else {
            t = X.temp %*% a.old / drop(crossprod(a.old))
            t = t / drop(sqrt(crossprod(t)))
        }

        if (na.Y) {
            u = Y.aux %*% b.old
            B = drop(b.old) %o% n.ones
            B[is.na.Y] = 0
            b.norm = crossprod(B)
            u = u / diag(b.norm)
            u = u / drop(sqrt(crossprod(u)))
        }
        else {
            u = Y.temp %*% b.old / drop(crossprod(b.old))
            u = u / drop(sqrt(crossprod(u)))
        }
        iter = 1

        #-- boucle jusqu'à convergence de a et de b --#
        repeat {
            if (na.X) a = t(X.aux) %*% u
            else a = t(X.temp) %*% u

              if (na.Y) b = t(Y.aux) %*% t
            else b = t(Y.temp) %*% t
```

M=X'Y.
a.old= loadings of X (p),
b.old=loadings of Y (q)

As previously imputed missing with NIPALS algorithm this code isn't needed

t=Xp / p'p scores of X

As previously imputed missing with NIPALS algorithm this code isn't needed

u=Yq/ q'q scores of Y

Calculates the loadings again from p=X'u and q=Y't

```
            if (nx != 0) {

#Lyn added in version 3: object loadord is created to rank the loadings before applying
the 0 to those not extracted
                loadord=a

#Lyn note: abs(a[order(abs(a))][nx]) returns the value (loading coefficient) of the
#corresponding max number to be extracted, i.e. extracting 20 vars, returns the 20th
coefficient.
                a = ifelse(abs(a) > abs(a[order(abs(a))][nx]),
                    (abs(a) - abs(a[order(abs(a))][nx])) * sign(a), 0)

# Lyn added in version 1: ranks the selected variables in order of importance and
outputs them in the object ordselvar#

                ordselvar[,h]=p-(rank(abs(loadord),ties="average"))+1

            }
            a = a / drop(crossprod(u))
            a = a / drop(sqrt(crossprod(a)))

            if (ny != 0) {
                b = ifelse(abs(b) > abs(b[order(abs(b))][ny]),
                    (abs(b) - abs(b[order(abs(b))][ny])) * sign(b), 0)
            }
            b = b / drop(crossprod(t))

            if (na.X) {
                t = X.aux %*% a
                A = drop(a) %o% n.ones
                A[is.na.X] = 0
                a.norm = crossprod(A)
                t = t / diag(a.norm)
                t = t / drop(sqrt(crossprod(t)))
            }
            else {
                t = X.temp %*% a / drop(crossprod(a))
                t = t / drop(sqrt(crossprod(t)))
            }

            if (na.Y) {
                u = Y.aux %*% b
                B = drop(b) %o% n.ones
                B[is.na.Y] = 0
                b.norm = crossprod(B)
                u = u / diag(b.norm)
                u = u / drop(sqrt(crossprod(u)))
            }
            else {
                u = Y.temp %*% b / drop(crossprod(b))
                u = u / drop(sqrt(crossprod(u)))
            }

            if (crossprod(a - a.old) < tol) break

            if (iter == max.iter) {

    warning(paste("Maximum number of iterations reached for the component", h),
                        call. = FALSE)
                break
            }

            a.old = a
            b.old = b
            iter = iter + 1
        }
```

Selects the top X variables to keep in the model based on the size of the p loadings

Selects the top Y variables to keep in the model based on the size of the q loadings

Recalculate the scores t and u based on the new loadings.

If the X loadings p are similar to the last iteration (per the Tol value) then the process is stopped.

```
   #-- deflation des matrices --#
    if (na.X) {
        X.aux = X.temp
        X.aux[is.na.X] = 0
        c = crossprod(X.aux, t)
        T = drop(t) %o% p.ones
        T[is.na.X] = 0
        t.norm = crossprod(T)
        c = c / diag(t.norm)
    }
    else {
        c = crossprod(X.temp, t) / drop(crossprod(t))
    }

    X.temp = X.temp - t %*% t(c)

    #-- mode canonique --#
    if (mode == "canonical") {
        if (na.Y) {
            Y.aux = Y.temp
            Y.aux[is.na.Y] = 0
            e = crossprod(Y.aux, u)
            U = drop(u) %o% q.ones
            U[is.na.Y] = 0
            u.norm = crossprod(U)
            e = e / diag(u.norm)
        }
        else {
            e = crossprod(Y.temp, u) / drop(crossprod(u))
        }

        Y.temp = Y.temp - u %*% t(e)
    }

    #-- mode regression --#
    if(mode == "regression") {
        if (na.Y) {
            Y.aux = Y.temp
            Y.aux[is.na.Y] = 0
            d = crossprod(Y.aux, t)
            T = drop(t) %o% q.ones
            T[is.na.Y] = 0
            t.norm = crossprod(T)
            d = d / diag(t.norm)
        }
        else {
            d = crossprod(Y.temp, t) / drop(crossprod(t))
        }

        Y.temp = Y.temp - t %*% t(d)
    }

    mat.t[, h] = t
    mat.u[, h] = u
    mat.a[, h] = a
    mat.b[, h] = b
    mat.c[, h] = c
    if (mode == "regression") mat.d[, h] = d

} #-- fin boucle sur h --#

#-- valeurs sortantes --#
rownames(mat.a) = rownames(mat.c) = X.names
rownames(mat.b) = Y.names
rownames(mat.t) = rownames(mat.u) = ind.names

dim = paste("comp", 1:ncomp)
colnames(mat.t) = colnames(mat.u) = dim
colnames(mat.a) = colnames(mat.b) = colnames(mat.c) = dim

cl = match.call()
cl[[1]] = as.name('spls')
```

The first component scores & loading are agreed and we deflate the X matrix using: Xnew = X – t c' where c=X't / t't

Not doing canonical modelling this section of code not used

The first component scores & loading are agreed and we deflate the Y matrix using: Ynew = Y – t d' where d=Y't / t't

```
    result = list(call = cl,
                  X = X, Y = Y, ncomp = ncomp, mode = mode,
                  keepX = keepX,
                  keepY = keepY,
                  mat.c = mat.c,
                  mat.t = mat.t,
                  variates = list(X = mat.t, Y = mat.u),
                  loadings = list(X = mat.a, Y = mat.b),
                  names = list(X = X.names, Y = Y.names, indiv = ind.names),

 # added by Lyn : version 1 if >1 component the order is the min selected in either
component.  selvarinord is output so can be used outside the function #
              selvarinord=apply(ordselvar,1,function(x) min(x))

      )
    if (length(nzv$Position > 0)) result$nzv = nzv

    class(result) = c("spls", "pls")
    return(invisible(result))
}

#############################################################################
#  MACRO 2: AMENDED valid FUNCTION renamed splscv                           #
#############################################################################
splscv <-
function(X,
         Y,
         ncomp = min(6, ncol(X)),
         method = "spls",
         mode = c("regression"),
         criterion = c("all"),
         keepX = NULL, keepY = NULL,
         validation = c("Mfold"),
         M = M,
         max.iter = 500,
         tol = 1e-06, ...)
{

    method = match.arg(method)

    #-------------------- sPLS ----------------------#
    if (any(c("spls") == method)) {

        #-- validation des arguments --#
        #-- do warning for mode + other warnings --#
        if (length(dim(X)) != 2)
            stop("'X' must be a numeric matrix.")

        mode = match.arg(mode)
        validation = match.arg(validation)

        X = as.matrix(X)
        Y = as.matrix(Y)

        n = nrow(X)
        q = ncol(Y)
        res = list()

        if (!is.numeric(X) || !is.numeric(Y))
            stop("'X' and/or 'Y' must be a numeric matrix.")

        if ((n != nrow(Y)))
            stop("unequal number of rows in 'X' and 'Y'.")

        if (any(is.na(X)) || any(is.na(Y)))
            stop("Missing data in 'X' and/or 'Y'. Use 'nipals' for dealing with NAs.")

# Added by Lyn: Version 2
# REMOVED the following check as variables with insufficient variation are removed on
# a variable by variable basis according to the cross folds used. See below:
#  We do not want them excluded from all analyses.
```

```
#             nzv = nearZeroVar(X, ...)
#        if (length(nzv$Position > 0)) {
#            warning("Zero- or near-zero variance predictors.
#  Reset predictors matrix to not near-zero variance predictors.
#  See $nzv for problematic predictors.")
#            X = X[, -nzv$Position]
#                res$nzv = nzv
#        }
          p = ncol(X)

        if (is.null(ncomp) || !is.numeric(ncomp) || ncomp <= 0 || ncomp > p)
            stop("Invalid number of components, 'ncomp'.")
        ncomp = round(ncomp)

        if (method == "spls") {
            if(is.null(keepX)) keepX = rep(ncol(X), ncomp)
            if(is.null(keepY)) keepY = rep(ncol(Y), ncomp)

            if (length(keepX) != ncomp)
                stop("length of 'keepX' must be equal to ", ncomp, ".")

            if (length(keepY) != ncomp)
                stop("length of 'keepY' must be equal to ", ncomp, ".")

            if (any(keepX > p))
                stop("each component of 'keepX' must be lower or equal than ", p, ".")

            if (any(keepY > q))
                stop("each component of 'keepY' must be lower or equal than ", q, ".")

        }

        #-- M fold  validation --#
        ##- define the folds
        if (validation == "Mfold") {
            if (is.null(M) | !is.numeric(M) | M < 2 | M > n)
                stop("Invalid number of folds, 'M'.")
            M = round(M)
            fold = split(sample(1:n), rep(1:M, length = n))
        }

        #-- compute MSEP and/or R2 --#
        if (any(criterion %in% c("all", "MSEP", "R2"))) {
            press.mat = Ypred = array(0, c(n, q, ncomp))
            MSEP = R2 = matrix(0, nrow = q, ncol = ncomp)
```

```
# Lyn added the following 3 rows in version 1 to create a blank matrices for filling in
```

```
            Svar=matrix(0, ncol(X), M)
            Ycorr=matrix(0, n, 2)
            ordSvar=matrix(0,p,M)

            for (i in 1:M) {
                omit = fold[[i]]
                X.train = X[-omit, ]
                Y.train = Y[-omit, ]
                X.test = matrix(X[omit, ], nrow = length(omit))
                  Y.test = matrix(Y[omit, ], nrow = length(omit))
```

```
## added by lyn: version 2 because there are SNPs which when folded have no variation
##  all 0's, then don't want process to fall down, so just exclude that variable
##  from that fold.  If a variable is predictive, then it will come up in all 50 runs
## but not in all folds as it has too low MAF with the sample size.   Lyn has
## programmed to include variables with at least 92% variation in the training set
## so minor alleles with <5% MAF won't be included
# Lyn added for version 3:  Minor code change from version 2:
# From sum(X>==0) to sum(x>=0 & x<=0.5).  This allows for 0'S or very small close to
zero # numbers derived from the NIPALS algorithm missing data imputation
# Otherwise macro still falls down with insufficient variation
```

```
          MAF <-apply(X.train[,1:ncol(X.train)], 2, function(x) sum(x>=0 & x<=0.5)) /
nrow(X.train[,1:ncol(X.train)]) * 100
           MAF2<-MAF>92  # put TRUE to those with >92% 0s.

###remove any with >92% 0s from the training & test X data ###
           X.train<-X.train[,!MAF2]
           X.test<-X.test[,!MAF2]


               X.train = scale(X.train, center = TRUE, scale = FALSE)
               xmns = attr(X.train, "scaled:center")

               Y.train = scale(Y.train, center = TRUE, scale = FALSE)
               ymns = attr(Y.train, "scaled:center")

               X.test = scale(X.test, center = xmns, scale = FALSE)

               #-- spls --#
               object = spls_lyn(X = X.train, Y = Y.train, ncomp = ncomp,
                                 mode = mode, max.iter = max.iter, tol = tol,
                                 keepX = keepX, keepY = keepY)
```

```
####################Lyn amended the code here : in version 1############
##    It will now export the predicted scores, the B coefficients     ##
#     and will list the variables selected                            ##
##    For each of the M models, for h components                      ##
##    It will also export the ordering of the selected variables      ##
#########################################################################
```

```
          predpar = predict(object,X.test)
          Y.hat=predpar$predict

          s.var=apply(abs(object$loadings$X),1,sum)>0
```

This tells us what is in the model and which variables are out

```
#  Lyn added version 2: As some variables excluded due to insufficient variation in
some #folds, the following code was added so that a complete list of variables in
created.
# Those not included as set to 0.
```

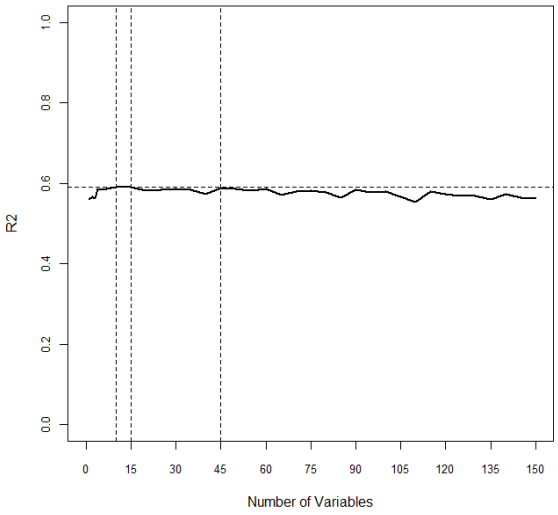
```
               MAF3<-cbind(MAF2,names(MAF2))
               colnames(MAF3)<-c("EIMOD","SNP")

               s.var2<-cbind(s.var,names(s.var))
               colnames(s.var2)<-c("CBMOD","SNP")

                s.var3 <- merge(s.var2,MAF3, by="SNP", all=TRUE)
                s.var3 [is.na(s.var3)] <- FALSE
                s.var5<-as.matrix(s.var3[,2])
                rownames(s.var5)<-s.var3[,1]
                 s.var6<- s.var5[order(row.names(s.var5)),]

                Svar[,i] <-as.matrix(s.var6) # changes the vector into a matrix
                rownames(Svar)=names(s.var6)
```

```
# Lyn amended here in Version 2: need to label selvarinord with var names.
# Merges on the variables which were not fitted above and get a complete list
# which all folds can be merged by.
                revar<-names(MAF2[!MAF2])
                selvarinord2<-cbind(object$selvarinord,revar)
                colnames(selvarinord2)<-c("CHORD","SNP")
                selvarinord3<-merge(selvarinord2, MAF3, by="SNP", all=TRUE)
                selvarinord4<-as.matrix(selvarinord3[,2])
                rownames(selvarinord4)<-selvarinord3[,1]
                 selvarinord5<- selvarinord4[order(row.names(selvarinord4)),]
                ordSvar[,i]=as.matrix(selvarinord5)
                rownames(ordSvar)=names(s.var6)

              for (h in 1:ncomp) {
             Y.mat = matrix(Y.hat[, , h], nrow = dim(Y.hat)[1], ncol= dim(Y.hat)[2])

             Y.hat[, , h] = sweep(Y.mat, 2, ymns, FUN = "+")
```

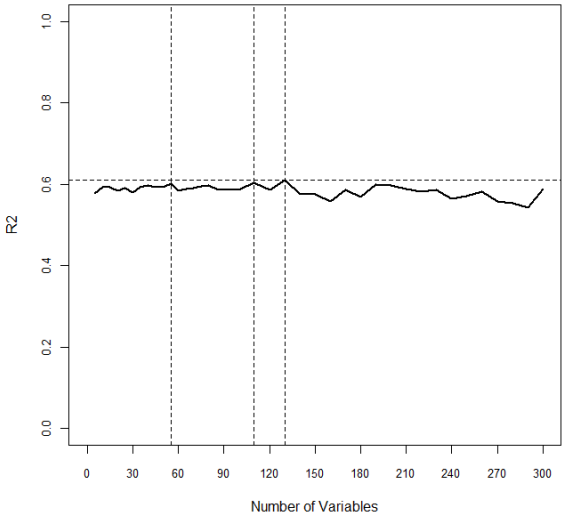
```
    # Added by lyn version1 : As modelling Larsen score, Yhat is only allowed to be
    #       0-160,  so limit results below

   Y.hat[, , h]<-apply(as.matrix(Y.hat[, , h]), 2 , function(x) ifelse(x<=0,0,
ifelse(x>=160,160,x)) )

                 press.mat[omit, , h] = (Y.test - Y.hat[, , h])^2
                 Ypred[omit, , h] = Y.hat[, , h]
                }

                Ycorr[omit,]<-cbind(Y.hat[, , ncomp], Y.test)   # added by Lyn
            } #end i

            for (h in 1:ncomp) {
                MSEP[, h] = apply(as.matrix(press.mat[, , h]), 2, mean, na.rm = TRUE)
                R2[, h] = diag(cor(Y, Ypred[, , h], use = "pairwise"))
            }
            colnames(MSEP) = colnames(R2) = paste('ncomp', c(1:ncomp), sep = " ")
            rownames(MSEP) = rownames(R2) = colnames(Y)

            if (q == 1) rownames(MSEP) = rownames(R2) = ""

            #-- valeurs sortantes --#
            if (any(criterion %in% c("all", "MSEP"))) res$MSEP = MSEP
             if (any(criterion %in% c("all", "R2"))) res$R2 = R2

##   3 lines added by lyn to ensure the new objects are output.
          res$Svar=Svar
          res$Ycorr=Ycorr
          res$ordSvar=ordSvar

        }

        #-- compute Q2 --#
        if (any(criterion %in% c("all", "Q2"))) {
            if (method == "pls") {
                Q2 = q2.pls(X, Y, ncomp, mode, M, fold, max.iter, tol)
            }
            else {
                Q2 = q2.spls(X, Y, ncomp, mode, keepX, keepY, M, fold, max.iter, tol)
            }

            Y.names = dimnames(Y)[[2]]
            if (is.null(Y.names)) Y.names = paste("Y", 1:q, sep = "")

            if (q > 1) {
                res$Q2$variables = t(Q2[, 1:q])
                res$Q2$total = Q2[, q + 1]
                rownames(res$Q2$variables) = Y.names
                colnames(res$Q2$variables) = paste('comp', 1:ncomp, sep = " ")
                names(res$Q2$total) = paste('comp', 1:ncomp, sep = " ")
            }
                     else {
                colnames(Q2) = ""
                rownames(Q2) = paste('comp', 1:ncomp, sep = " ")
                res$Q2 = t(Q2)
            }
        }
    }

    method = paste(method, "mthd", sep = ".")
    class(res) = c("valid", method)
    return(invisible(res))
}
```

Appendix E: Graphs determining optimum number of variables to extract based on R^2-CV derived for each of the 40 blocks of data.

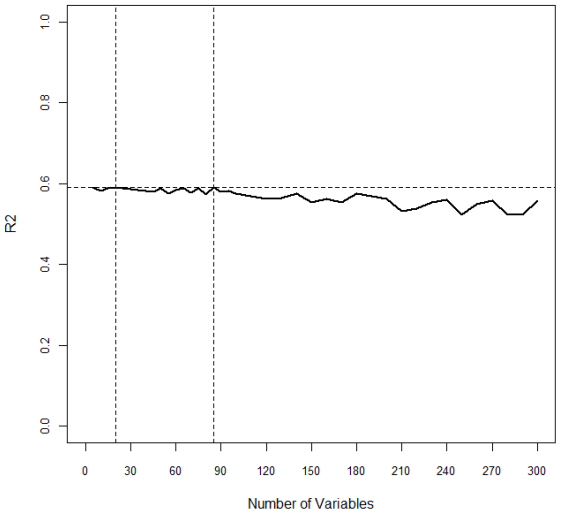
Chromosome 1, part 1 (SNP=8257)
The maximum R^2 of 0.590 is obtained when 10 variables are extracted. However, a very similar R^2 is found at 15 (R^2=0.589) and 45 (R^2=0.588). After this point the R^2 gradually decreases. To allow the maximum chance of selecting variables for final higher level model which may be important, it was decided to extract 45 variables from this model.



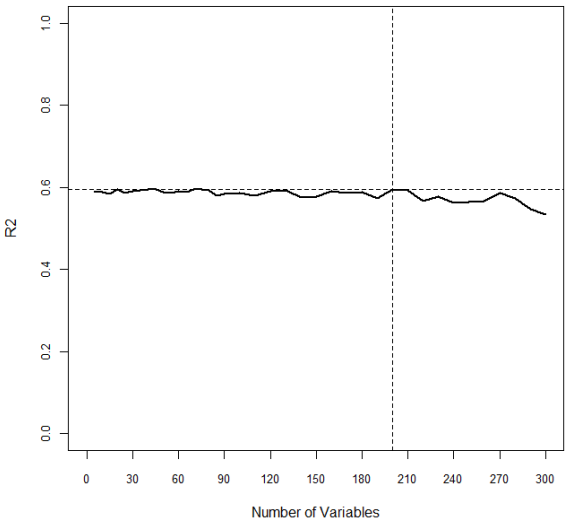
Chromosome 1, part 2 (SNP=8258)
The maximum R^2 of 0.6096655 is obtained when 130 variables are extracted. From the graph, it appears to gradually increase up to this point (almost reaching it at 55 and 110) and then decreases after this point. Therefore, it was decided to extract 130 variables from this model.



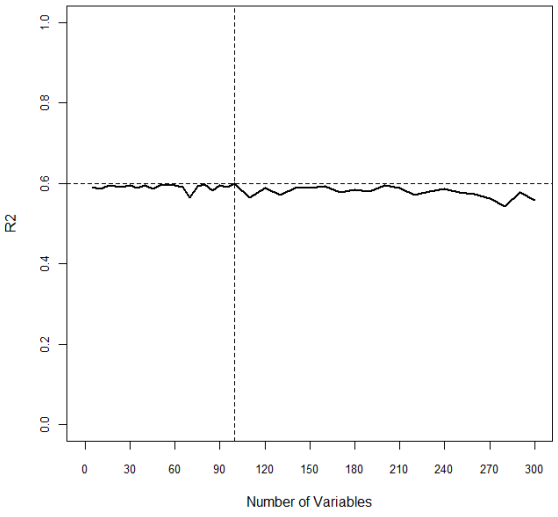
Chromosome 1, part 3 (SNP=8194)
The maximum R^2 of 0.5907745 is obtained when 85 variables are extracted. However, this value is almost obtained with just 20 variables. As the graph is relatively stable at the 0.59 level up to 85 variables, it was decided to extract 85 variables from this model.



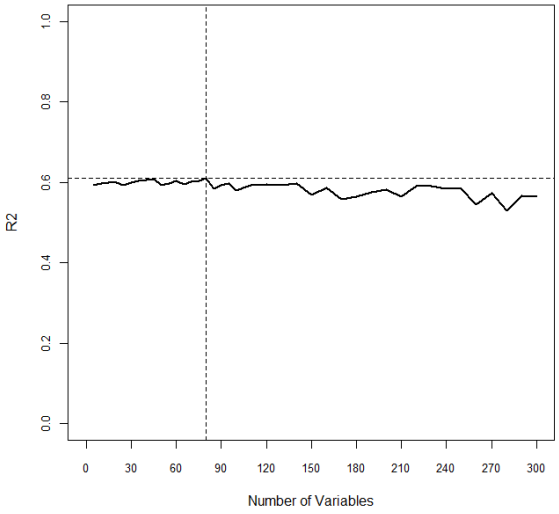
Chromosome 2, part 1 (SNP=8884)
The maximum R^2 of 0.5953790 is obtained when 200 variables are extracted. The graph is relatively stable up to 200 variables and then starts to decrease after this point. Therefore it was decided to extract 200 variables from this model.



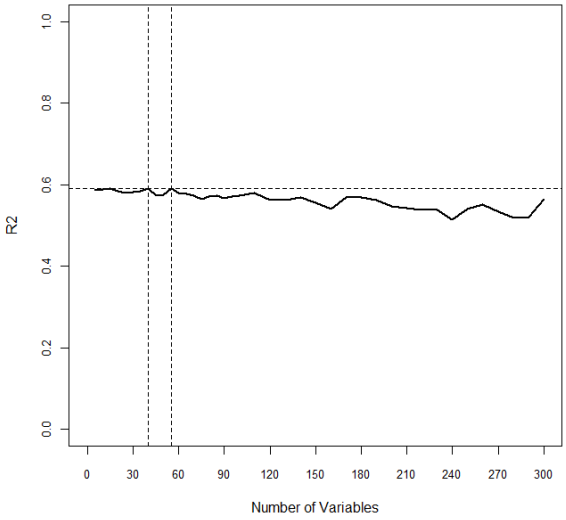
Chromosome 2, part 2 (SNP=8872)
The maximum R^2 of 0.5989761 is obtained when 100 variables are extracted. The graph is relatively stable up to 100 variables and then starts to decrease after this point. Ttherefore it was decided to extract 100 variables from this model.



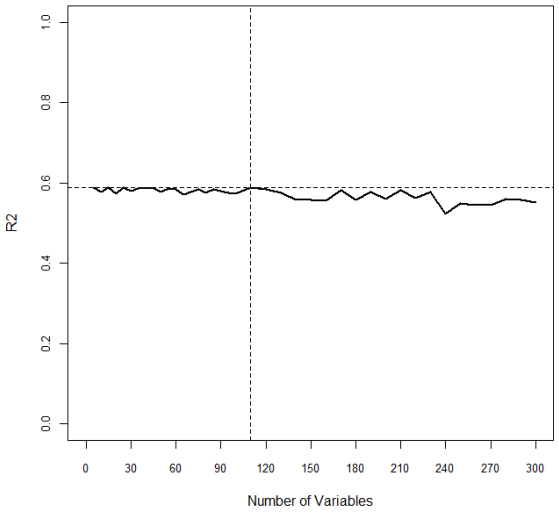
Chromosome 2, part 3 (SNP=8796)
The maximum R^2 of 0.6104767 is obtained when 80 variables are extracted. The graph is relatively stable up to 80 variables and then starts to decrease after this point. Therefore it was decided to extract 80 variables from this model.



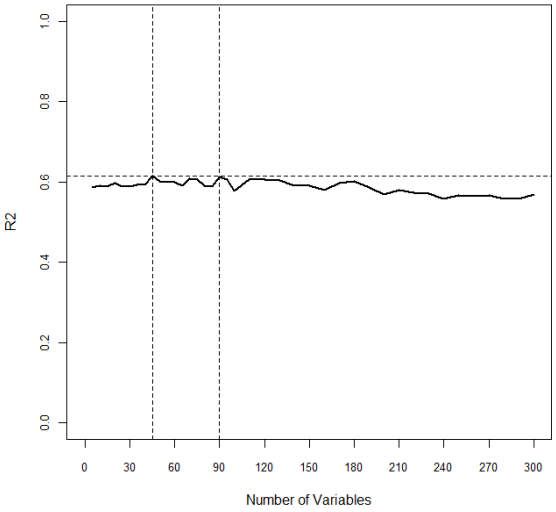
Chromosome 3, part 1 (SNP=11361)
The maximum was observed at 40 variables however it was almost reached again at 55 variables (R^2=0.5913 and 0.5907 respectively). The R2 was relatively stable prior to 40 but we observed at dip at 50 variables before it increased at 55 variables. It was therefore decided to extract 40 variables for this model.



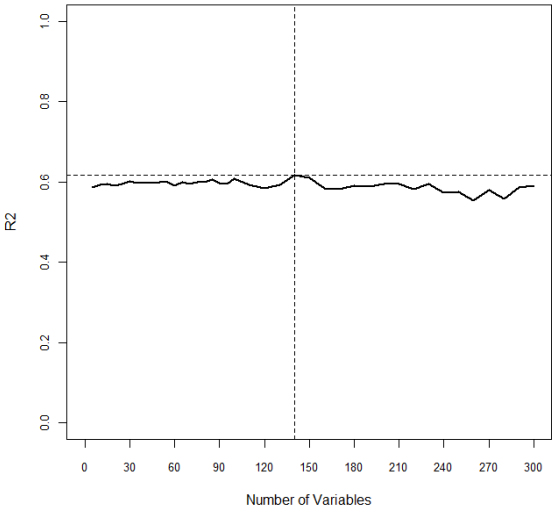
Chromosome 3, part 2 (SNP=11298)
The maximum was observed at 110 variables (R^2=0.589). It is stable up to this point but decreases after hence 110 variables will be extracted for this model.



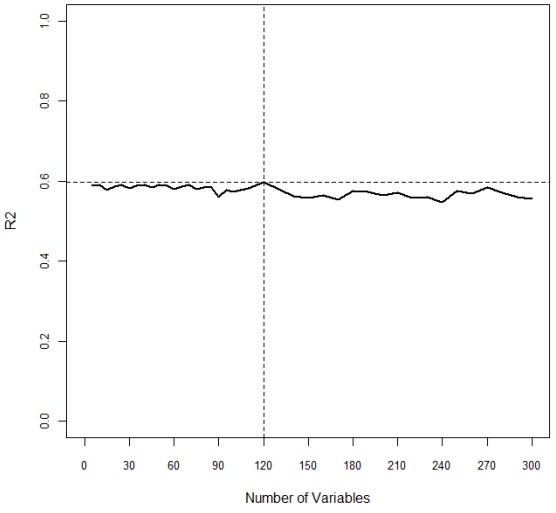
Chromosome 4, part 1 (SNP=10142)
The maximum was observed at 45 R^2= 0.6151818, then almost again at 90, R^2 =0.6113111. It is very variable between 45 and 120 when it almost reaches the same level again. It was decided to pick a midpoint in order to give variables maximum opportunity for being selected. Therefore 90 variables were chosen to be extracted for this model.



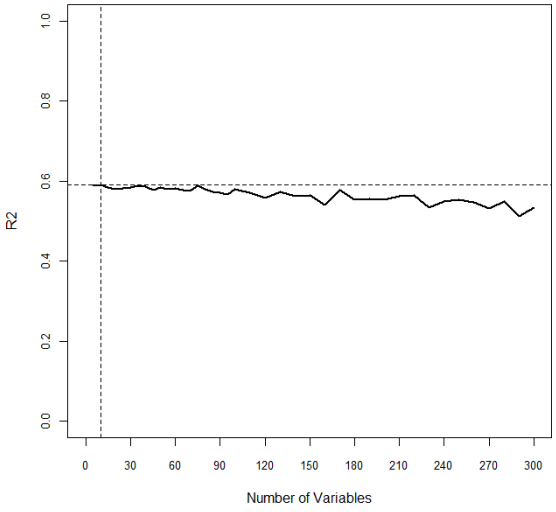
Chromosome 4, part 2 (SNP=10042)
This is more like the plot expected on each set of data. The line steadily increases until a maximum observed at 140, R^2= 0.6175557. After which, the R^2 decreases steadily. Therefore 140 variables were extracted for this model.



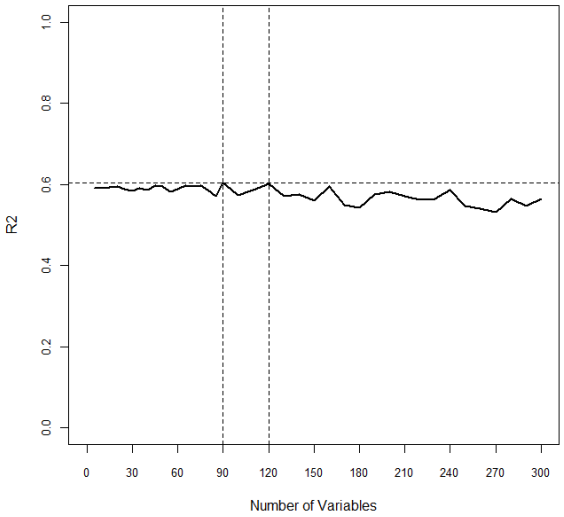
Chromosome 5, part 1 (SNP=10077)
The maximum R^2 is observed at 120, R^2= 0.5972783. There is quite a bit of variability after 80 however to give variables the best chance of being included, 120 variables were extracted for this model.



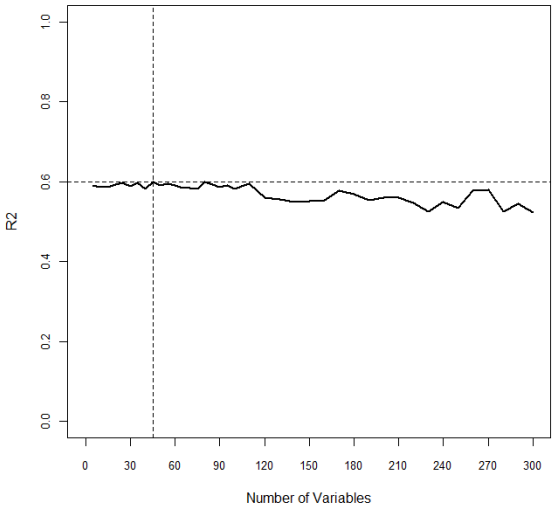
Chromosome 5, part 2 (SNP=9989)
This plot is interesting as the maximum is observed at just 10 variables extracted (R^2=0.5910226) and the line decreases from that point onwards. It was decided to extract 10 variables for this model as it is assumed that the addition of more variables does not aid in the prediction of the Larsen score.



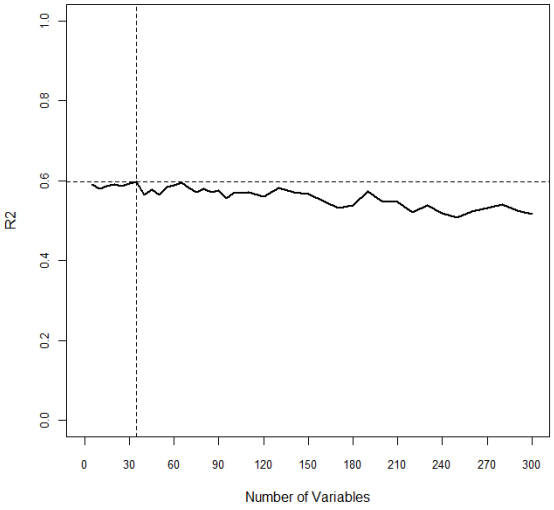
Chromosome 6, part 1 (SNP=8396)
This plot is very variable and hard to determine the optimum number of variables. It is clear however that the plot steadily decreases after 120 variables. The maximum is observed at 90 variables (R^2=0.6030127) and a similar R^2 at 120 variables R2=0.6011744. Given the instability before 90 variables, extracting 90 variables was selected for this model.



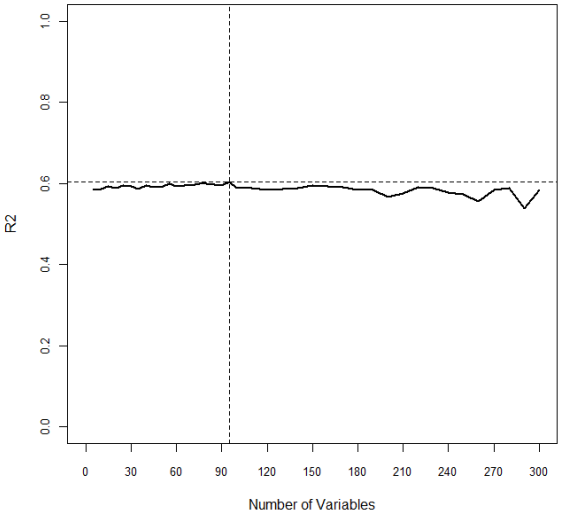
Chromosome 6, part 2 (SNP=8325)
Although the maximum is observed at 45 variables 0.5993820, approximately this level is retained until 110 variables (R2=0.5943365). To give variable the maximum opportunity to be selected, 110 variables were decided to be extracted for this model.



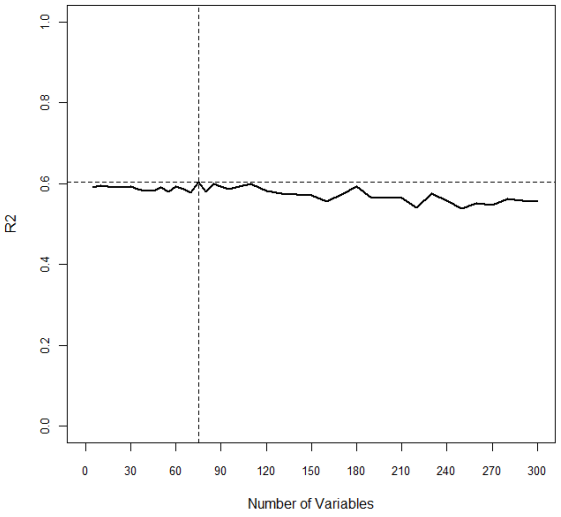
Chromosome 6, part 3 (SNP=8135)
The maximum is observed at just 35 variables R^2=0.5972813. However, there is another peak at 65 variables which results in nearly the same R^2 (R^2=0.5956908). After the peak at 65 the line steadily decreases, therefore 65 was chosen as the number of variables to extract from this model.



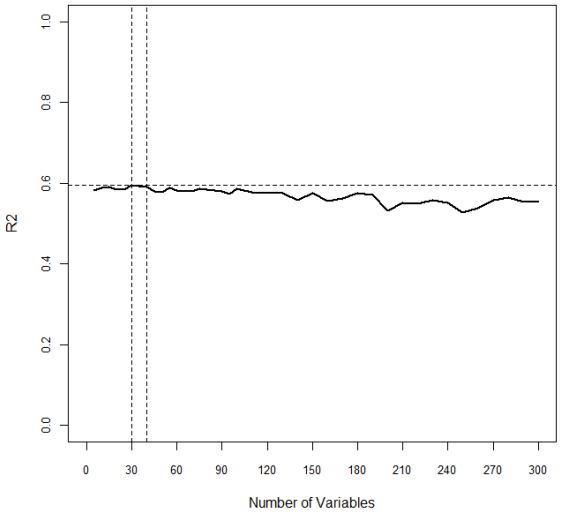
Chromosome 7, part 1 (SNP=8945)
The graph steadily increases to the point of 95 variables (R^2=0.6025330) after which point it decreases and becomes more variable. Therefore extracting 95 variables was selected for this model.



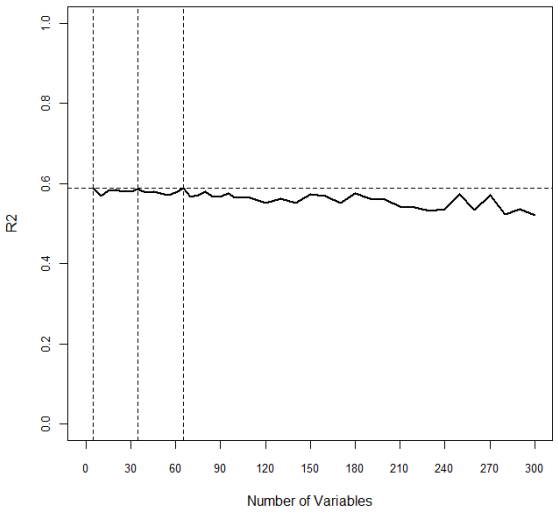
Chromosome 7, part 2 (SNP=8859)
The graph is quite varied but does maximise at 75 variables (R^2=0.6027475). Although there are two small peaks after that it does generally decrease. Therefore extracting 75 variables was selected for this model.



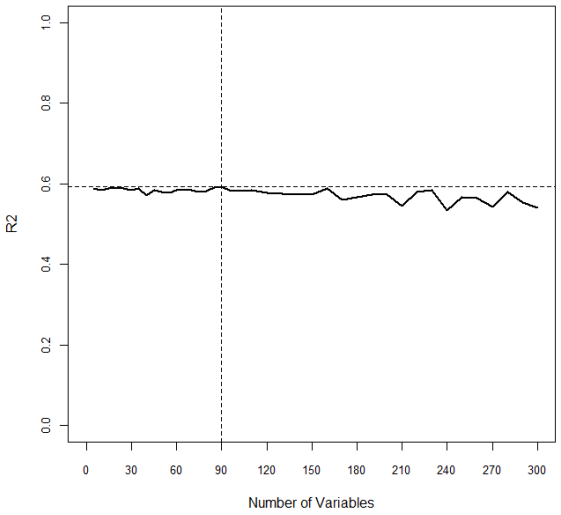
Chromosome 8, part 1 (SNP=9393)
The maximum was found at 30 variables (R^2=0.5936440) however was almost the same at 40 variables before it steadily decreased. Therefore 40 variables were selected for extraction for this model.



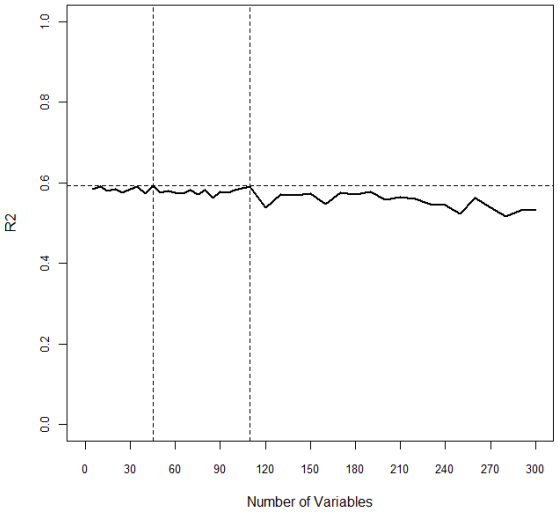
Chromosome 8, part 2 (SNP=9345)
The maximum observed at just 5 variables (R^2=0.5885593) however it is relatively consistently high until 65 variables (R^2=0.5872437) after which it steadily decreases. Therefore 65 variables were selected for extraction for this model.



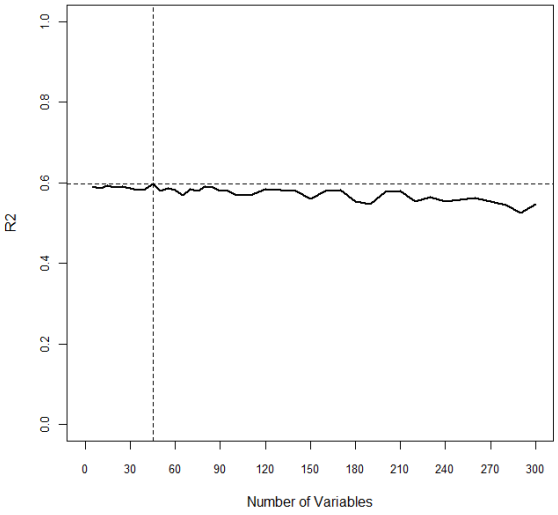
Chromosome 9, part 1 (SNP=8126)
The maximum observed at 90 variables (R^2=0.5920790) after which point the R2 decreases. Therefore 90 variables were selected for extraction for this model.



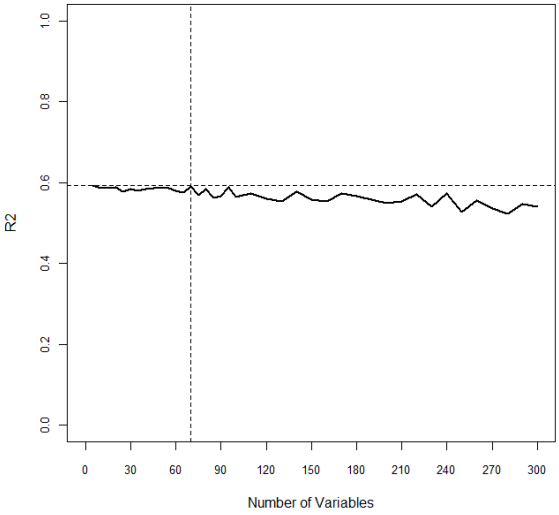
Chromosome 9, part 2 (SNP=8089)
The maximum observed at 45 variables (R^2=0.5929606) however the R^2 remains high up to 110 variables after which point it decreases. To maximise the opportunity for variables to be selected 110 variables were selected for extraction for this model.



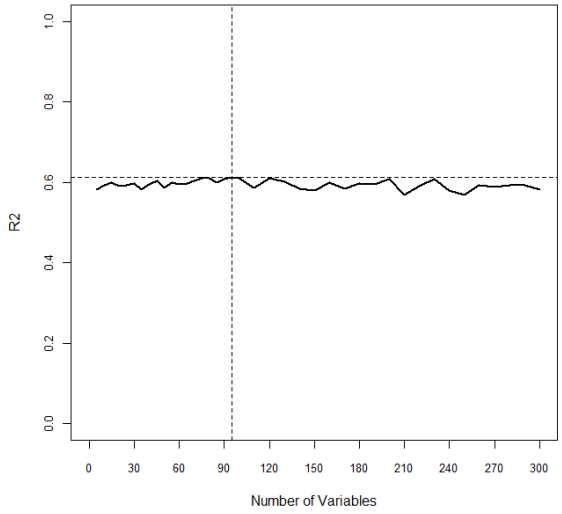
Chromosome 10, part 1 (SNP=8257)
The maximum R^2 of 0.5969518 occurs at 45 variables after which point the R^2 decreases. Therefore 45 variables will be extracted for this model.



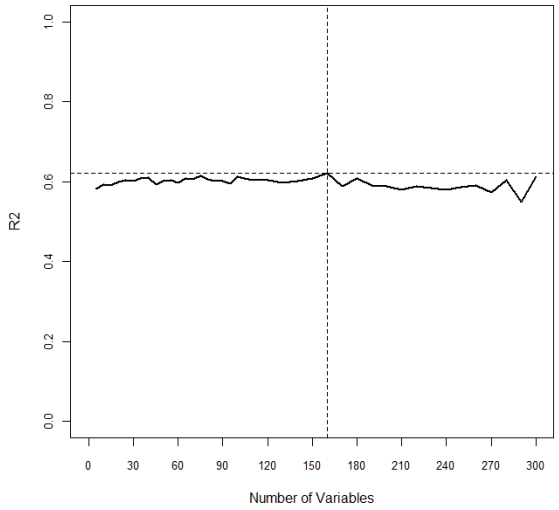
Chromosome 10, part 2 (SNP=8162)
The maximum R^2 of 0.5926617 occurs at 70 variables after which point the R^2 decreases. Therefore 70 variables will be extracted for this model.



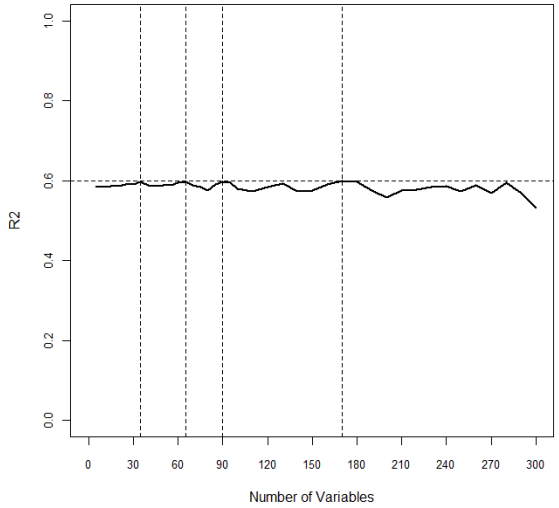
Chromosome 11, part 1 (SNP=7742)
The R^2 increases up to 95 variables after which is becomes more unstable. Although there are some peaks which almost match the maximum at 200 and 230 variables it was decided extracting 95 variables would be sufficient given the variation in R^2 after that point.



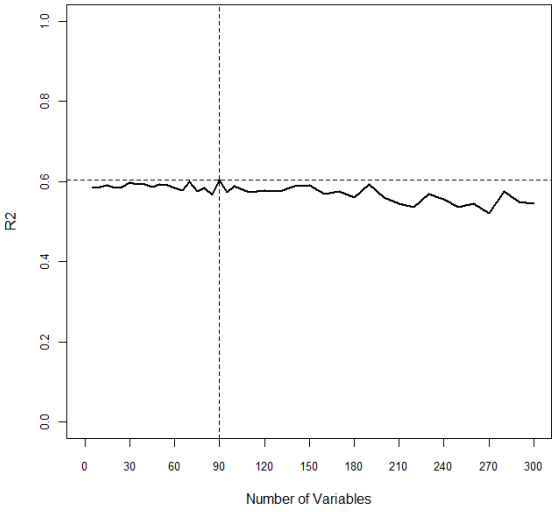
Chromosome 11, part 2 (SNP=7679)
The R^2 increases up to 160 variables (R^2=0.6202291) after which it decreases. Therefore 160 variables were chosen to extract for this model.



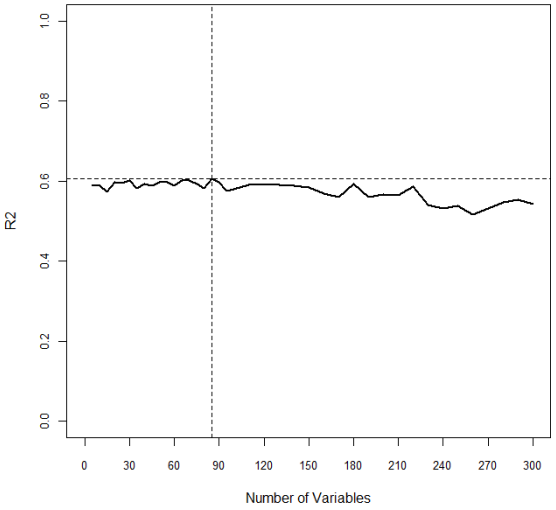
Chromosome 12, part 1 (SNP=7837)
This graph is harder to determine the optimum number of variables to extract as although the maximum R^2 is observed at 170 (R^2=0.6001613), it is almost at the same level at 35, 65, 90 and 180 variables. As this is a preliminary stage and variables can be removed at the upper level model fitting, it was decided to extract 170 with the provision that some extracted may not be very predictive.



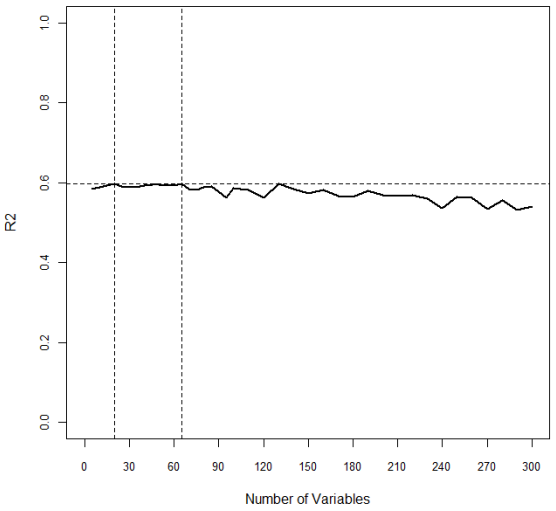
Chromosome 12, part 2 (SNP=7765)
The maximum R^2 was observed at 90 variables (R^2=0.6037449) after which it decreased. Although there is some variability before 90 variables in the R^2 estimates, 90 variables were chosen to be extracted.



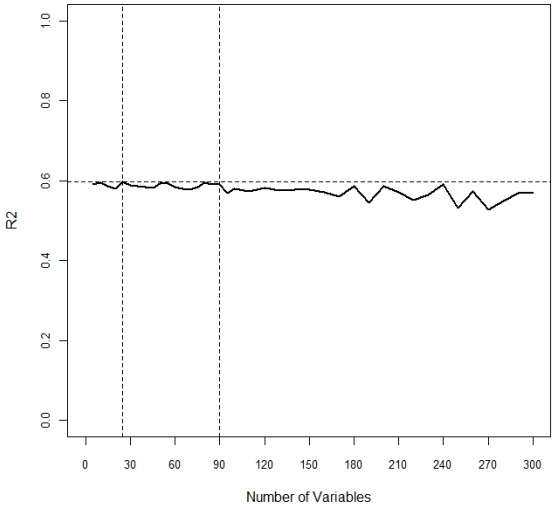
Chromosome 13, part 1 (SNP=6153)
Although slightly variable, the maximum R^2 was observed at 85 variables (R^2=0.60634). As there was a clear decrease after this point, 85 was selected as the number of variables to export.



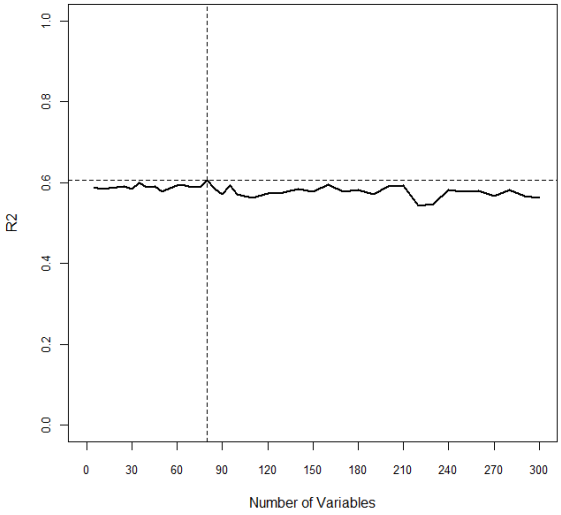
Chromosome 13, part 2 (SNP=6093)
The maximum R^2 was observed at 20 variables (R^2=0.5975580) however it was almost as high again at 65 variables (R^2=0.5974014), after which the R^2 decreases. Therefore 65 variables were chosen to be extracted.



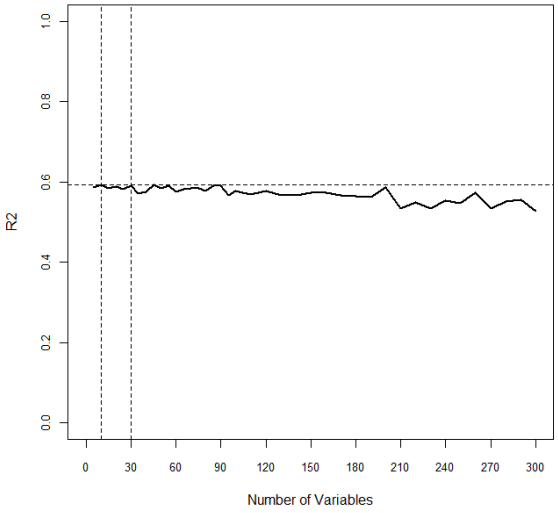
Chromosome 14, (SNP=10279)
The maximum R^2 of 0.5967481 was observed at 25 variables however it remains a a similar level up to 90 variables 0.5914446. It was therefore decided to export 90 variables for this model.



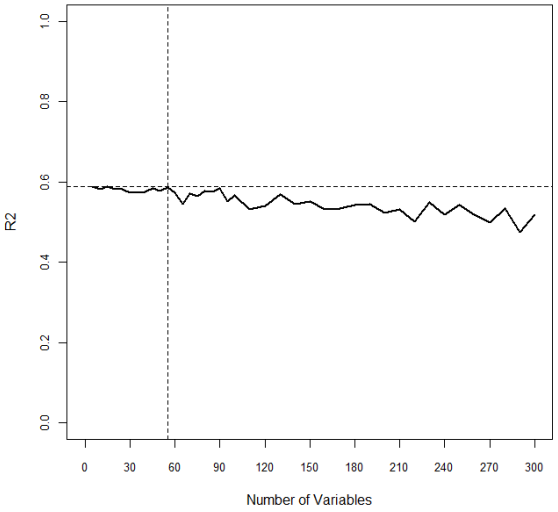
Chromosome 15, (SNP=9247)
The maximum R^2 was at 80 variables (R^2=0.604956) and as it increases before this point and decreases after this point, 80 variables were selected for extraction.



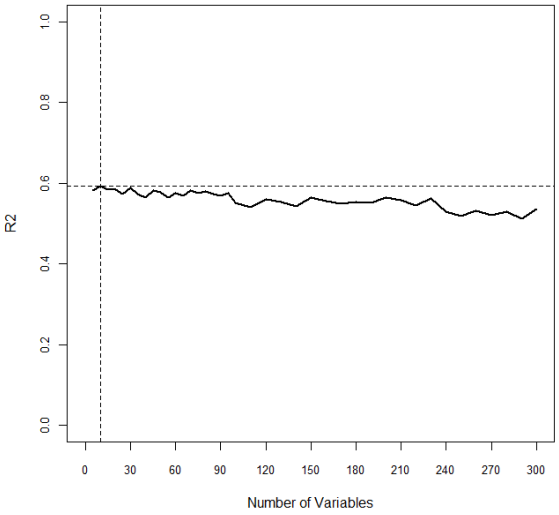
Chromosome 16, (SNP=9232)
Although the maximum was observed at 10 variables (R^2=0.5932914), it remains approximately level until 30 variables (R^2=0.5909528) where it becomes more unstable with a decreasing trend. Therefore 30 variables were selected to be extracted.



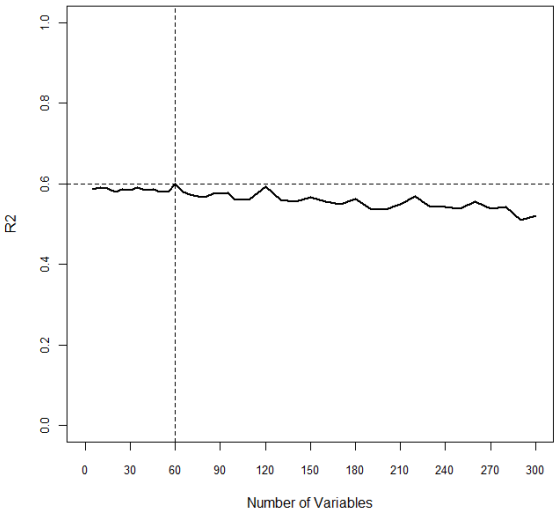
Chromosome 17, (SNP=8672)
Although the highest R^2 is observed at just 5 variables (R^2=0.5889406), the estimate is relatively stable & almost as high at 55 variables (R^2=0.5856562). After 55 the estimate is more variable and it decreases. Therefore 55 variables were chosen to be extracted for this model.



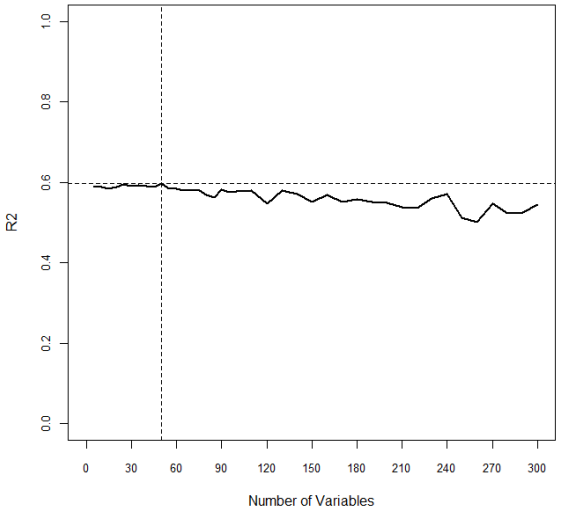
Chromosome 18, (SNP=10629)
Although the maximum R^2 is at 10 variables (0.5926424), the same value is almost reached at 30 variables (R^2=0.5887089). In order to export the maximum predictive SNPs possible (and there is opportunity to exclude variables at a later date if found to be not predictive), 30 variables were selected to be exported.



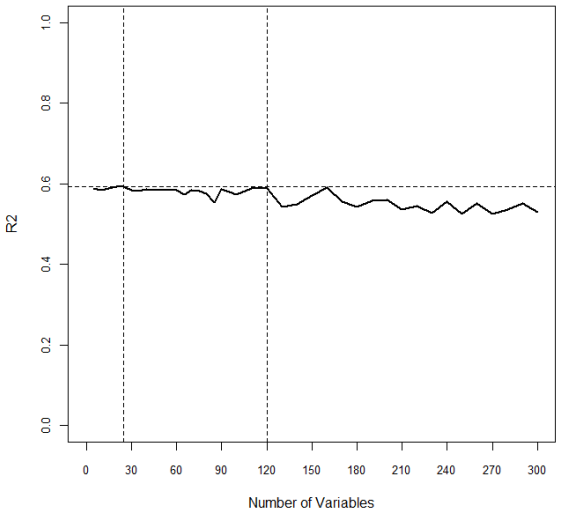
Chromosome 19, (SNP=6277)
The maximum R^2 is observed at 60 variables (R^2=0.5999831) after which is generally decreases (except for a spike at 120). It was decided to extract 60 variables for this model.



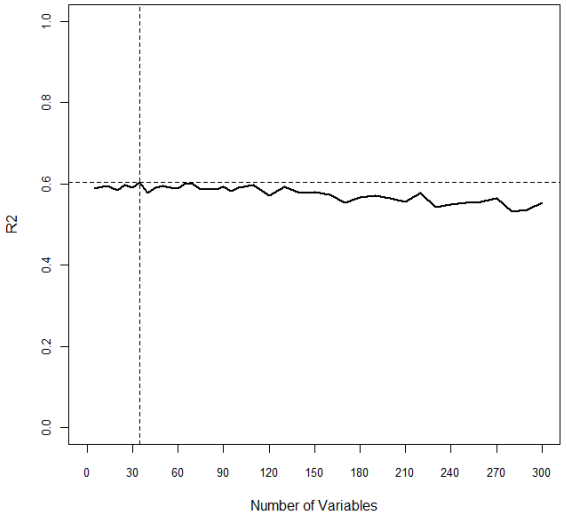
Chromosome 20, (SNP=7951)
The maximum R^2 is observed at 50 variables (R^2=0.5968821). Prior to this point the estimate is relative stable and after this point R^2 gradually decreases. Therefore 50 variables were chosen for extraction.



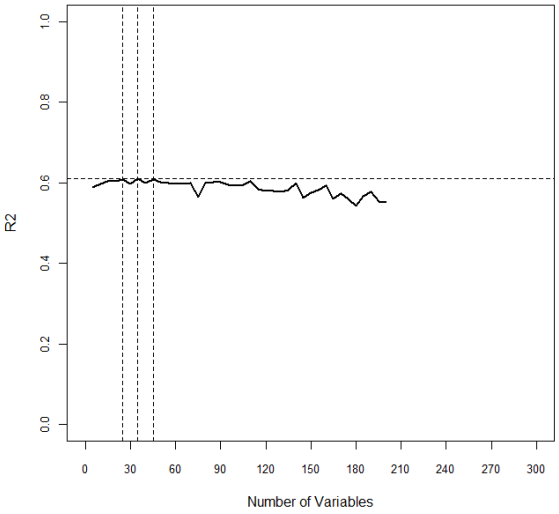
Chromosome 21, (SNP=5532)
Although the maximum R^2 was observed at 25 variables (R^2= 0.5926111), it remains at a similar level with just small variation until 120 variables (R^2=0.5879888). After 120 we see a decrease and therefore 120 was chosen as the number of variables to extract for this model.



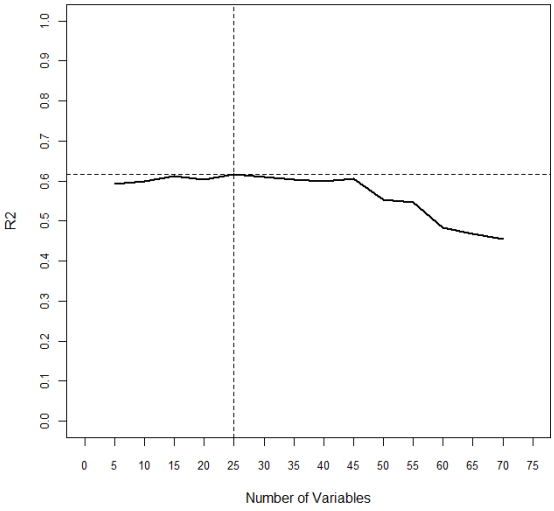
Chromosome 22, (SNP=5622)
The maximum R^2 is observed at 35 variables (0.6028986) after which the R^2 declines. Therefore 35 was chosen as the number to export.



X chromosome, (SNP=499)
The maximum is observed at 35 (0.6094233) but it is almost reached again at 45 and 25. In order to export as many as possible important variables it was decided to export 45 variables.



XY Pseudo autosomal region of X, (SNP=74)
As there are only 74 variables to choose from, the region of extracting from 5 to 70 variables was examined. The maximum R^2 was observed at 25 variables with a decrease in R^2 observed after this. Therefore 25 variables were selected for extracting in this model.



Appendix F: Code to produce GWAS analysis

```
#-------------------------------------------------------------------------#
# 2012: Program fits SPLS analysis for the larsen score                   #
# Analysis 3: GWAS MODELLING                                              #
# This program performs the cross validation with imputed data            #
# This program is similar to the original GWAS modelling  except         #
# The mixomics macro has been amended to output the order of all SNPS     #
# rather than SNPS > number extracted being ranked equal last             #
# in addition, rather than exploring optimum number of SNPs to export for#
# each chromosome, 200 are extracted and then hoped the excess be removed#
# in the higher level model                                               #
#-------------------------------------------------------------------------#
require(mixOmics)
require(gtools)

#### read in the macros I've adapted from mixomics ##############
source("D:\\Lyns Stuff\\PHD\\R with Gora Data\\R functions code from MixOmics
3_0\\Macros adapted from Mixomics_version 4.R")

### amend the location here and it will follow through all code below ###
##data location###
datloc<-"D:\\Lyns Stuff\\PHD\\3rd year plan and record of work\\GWAS modelling\\"
##output location ##
outloc<-"D:\\Lyns Stuff\\PHD\\3rd year plan and record of work\\GWAS
modelling\\run10folddata\\"

#####################################################################
# The following macros runs through the process of the lower level #
# modelling, each time selecting 200 variables in the PLS          #
# however the ordering of variables is continuous from 1 to max SNP#
# the n times selected though is based on being in the top 200     #
#####################################################################

# inport required Y data and ensure only GWAS subjects kept
yin <-paste(datloc,"mody.csv", sep="")
mody <- read.csv(yin, sep="," , header=T)
gwin <-paste(datloc,"GWAS_subj.csv", sep="")
gwas <- read.csv(gwin, sep="," , header=T)
rownames(gwas)=gwas$subjid
rownames(mody)=mody$subjid
larsen <-merge(gwas, mody, all.x = TRUE)
larsen <-as.numeric(larsen[,11])


#############################################################################
#This macro reads each block of snps per chromosomes,keeps only the chosen       #
# variables  with an average order of selection less than 200                 #
#############################################################################
rodat1 <- defmacro(chrid,
                   expr={

   indata <-paste(datloc,chrid,".csv", sep="")
   datx <- read.csv(indata, sep="," , header=T)

   inkeep <-paste(outloc,chrid,"final.csv", sep="")
   keepx <- read.csv(inkeep, sep="," , header=T)
   selvar<-keepx[which(keepx$aordSvar<200),2]
   myvars <- names(datx) %in% selvar
   chrid<- datx[,myvars]

})


rodat1(chrid="chr1a")
rodat1(chrid="chr1b")
rodat1(chrid="chr1c")
rodat1(chrid="chr2a")
rodat1(chrid="chr2b")
rodat1(chrid="chr2c")
rodat1(chrid="chr3a")
```

```
rodat1(chrid="chr3b")
rodat1(chrid="chr4a")
rodat1(chrid="chr4b")
rodat1(chrid="chr5a")
rodat1(chrid="chr5b")
rodat1(chrid="chr6a")
rodat1(chrid="chr6b")
rodat1(chrid="chr6c")
rodat1(chrid="chr7a")

rodat1(chrid="chr8a")
rodat1(chrid="chr8b")
rodat1(chrid="chr9a")
rodat1(chrid="chr9b")
rodat1(chrid="chr10a")
rodat1(chrid="chr10b")
rodat1(chrid="chr11a")
rodat1(chrid="chr11b")
rodat1(chrid="chr12a")
rodat1(chrid="chr12b")
rodat1(chrid="chr13a")
rodat1(chrid="chr13b")

rodat1(chrid="chr15")
rodat1(chrid="chr16")
rodat1(chrid="chr17")
rodat1(chrid="chr18")
rodat1(chrid="chr19")
rodat1(chrid="chr20")
rodat1(chrid="chr21")
rodat1(chrid="chr22")

rodat1(chrid="chr7b")
rodat1(chrid="chr14")
rodat1(chrid="chr23")

###now need to merge all the columns together to create 1 X dataset ###

fxvars<-
cbind(chr1a,chr1b,chr1c,chr2a,chr2b,chr2c,chr3a,chr3b,chr4a,chr4b,chr5a,chr5b,chr6a,chr
6b,chr6c,chr7a,chr7b,chr8a,chr8b, chr9a,chr9b,chr10a,chr10b,
chr11a,chr11b,chr12a,chr12b, chr13a,chr13b,chr14,chr15,chr16, chr17,chr18,chr19,chr20,
chr21,chr22,chr23)
sfxvars<-unique(names(fxvars))
finxv<- fxvars[,sfxvars]
ncol(finxv)

names(finxv)

#outfdat <-paste(outloc,"NIPIMPupperlevelmodel_mixomics.csv", sep="")
#write.csv(finxv,outfdat)

#########################################################################
##########keep top 100 variables                                 #########
#########################################################################

###- decrease to 100 vars and see what happens to prediction ****;

   inkeep <-paste(outloc,"upperlevelmodel_final_selvarsfinal.csv", sep="")
   keepx <- read.csv(inkeep, sep=","  , header=T)
   selvar<-keepx[which(keepx$X<101),2]
   myvars <- names(finxv) %in% selvar
   xmodel<- finxv[,myvars]
names(xmodel)
#outfdat <-paste(outloc,"top100variables.csv", sep="")
#write.csv(xmodel,outfdat)

   res <-pls(xmodel, larsen, ncomp=1, max.iter=500, tol=1e-09)  #spls or pls gets same
model here#
   pred <-predict(res, xmodel)

       # Any predicted values <0 or >160 amended to be predicted as 0 and 160 #
```

```
       comp2<-apply(as.matrix(pred$predict[,1,1]), 2 , function(x) ifelse(x<=0,0,
ifelse(x>=160,160,x)) )
       predvals<-data.frame(cbind(comp2,larsen))

       #plot actual value vs predicted value #
outgraph<-paste(outloc,"upperlevelmodelfinal100variables_101",".tif",sep="")
tiff(outgraph, height=1200, width=2400, res=600, units="px", pointsize=6, compression =
"lzw")
par(mar=c(5.1, 5.1, 3.1, 2.1))
       plot (predvals$V1,predvals$larsen, xlab="Predicted Larsen Score", ylab="Larsen
score",
            pch=4, cex=0.8, cex.axis=0.6,
          xlim=c(0,180), xaxp=c(0,180,9), ylim=c(0,180), yaxp=c(0,180,9))
       #   Get the correlation of the plot #
       corrv<-round(cor(predvals$V1,predvals$larsen),digits=3)
       text(120,5,"r=")
       text(140,5,corrv)
       abline(1,1)

dev.off()

# calculate the difference between predicted & actual larsen score
       preddiff2<-abs(predvals$larsen-predvals$V1)
       mean(preddiff2,na.rm=T)
       median(preddiff2,na.rm=T)
       sd(preddiff2,na.rm=T)
       min(preddiff2,na.rm=T)
       max(preddiff2,na.rm=T)
       sum(is.na(preddiff2))


       #**********have a look at % we get correct if predicting larsen >5 ***;
       act_grp <- cut(predvals$larsen, breaks=c(-0.1,5,20,60,164))
       pre_grp <- cut(predvals$V1, breaks=c(-0.1,5,20,60,164))
       table(act_grp,pre_grp)



# ***  under cross validation ######

cv_model<- valid(finaldat, larsen, ncomp=2, method="pls", mode="regression",
criterion="all",
         validation="Mfold", M=5, max.iter=500, tol=1e-09
          )
cv_model

old<-read.csv("D:\\Lyns Stuff\\PHD\\2nd year plan and record of work\\detailed write up
of PLS larsen score GWAS analyses\\mindata.csv", sep=",", header=T)
names(old[,2:505])
names(finaldat)


##############################################################################
##if you want to get the parameter estimates of the model run the following:####
##############################################################################


   indata <-paste(outloc,"upperlevelmodel_final_selvarsfinal.csv", sep="")
   resord <- read.csv(indata, sep="," , header=T)
   resord2<-resord[1:100,2]  # 1:505 uses top 500 SNP & 5 env,

   indat <-paste(datloc,"upperlevelmodel_final_selvars.csv", sep="")
   datx <- read.csv(indat, sep="," , header=T)
   myvars <- names(datx) %in% resord2
   finaldat<- datx[,myvars]

   spls.final <-pls(finaldat, larsen, ncomp=1, max.iter=500, tol=1e-09)  #spls or pls
gets same model here#
   larsen.predict <-predict(spls.final, finaldat)

   ##output variables from the prediction model -beta hat are the coefficients of the
regression###
```

```
   # NOTE: coefficient below (1) represents 1 component only, increase if more
components #
   vpred<-cbind(names(finaldat),round(larsen.predict$B.hat[,,1], digits=3))
   cnames<-c("xname","p_est")
   colnames(vpred)<-cnames
  vpred

   ##############################################################
   #  Because the code above doesn't output the intercept run #
   # the following copied from Mixomics predict which will    #
   ##############################################################

    newdata=finaldat
   X = spls.final$X
   Y = spls.final$Y
   q = ncol(Y)
   p = ncol(X)

    #-- initialisation des matrices --#
    ncomp = spls.final$ncomp
    a = spls.final$loadings$X
    b = spls.final$loadings$Y
    c = spls.final$mat.c

    means.X = attr(X, "scaled:center")
    means.Y = attr(Y, "scaled:center")
    sigma.X = attr(X, "scaled:scale")
    sigma.Y = attr(Y, "scaled:scale")

    newdata = as.matrix(newdata)
    ones = matrix(rep(1, nrow(newdata)), ncol = 1)
    ##- coeff de regression
    B.hat = array(0, dim = c(p, q, ncomp))
    ##- prediction
    Y.hat = array(0, dim = c(nrow(newdata), q, ncomp))
    ##- variates
    t.pred = array(0, dim = c(nrow(newdata), ncomp))

    #-- calcul de la prediction --#
    for(h in 1:ncomp){
           W = a[, 1:h] %*% solve(t(c[, 1:h]) %*% a[, 1:h])
           B = W %*% drop(t(b[, 1:h]))
           B = scale(B, center = FALSE, scale = 1 / sigma.Y)
           B = as.matrix(scale(t(B), center = FALSE, scale = sigma.X))
           intercept = -scale(B, center = FALSE, scale = 1 / means.X)
           intercept = matrix(apply(intercept, 1, sum) + means.Y, nrow = 1)
           Y.hat[, , h] = newdata %*% t(B) + ones %*% intercept
           t.pred[, h] = scale(newdata, center = means.X, scale = sigma.X) %*% W[, h]
              B.hat[, , h] = B
    }  #end h

   intercept


})
```

Appendix G: Code to produce GWAS analysis using 80% of data for training and 20% of the data to independently test the model

```
#-----------------------------------------------------------------------#
# 2013: Program fits SPLS analysis for the larsen score                 #
# Analysis 3: GWAS MODELLING                                            #
# This program performs the cross validation  with imputed data         #
# This program is similar to the original GWAS modelling  except        #
# The mixomics macro has been amended to output the order of all SNPS   #
# rather than SNPS > number extracted being ranked equal last           #
# in addition, rather than exploring optimum number of SNPs to export for#
# each chromosome, 200 are extracted and then hoped the excess be removed#
# in the higher level model                                             #
#                                                                       #
#  This program only uses 80% of the data to form the prediction model  #
#  The remaining 20% is used as an independent prediction sample        #
#-----------------------------------------------------------------------#

require(mixOmics)
require(gtools)

#### read in the macros I've adapted from mixomics ##############
source("D:\\Lyns Stuff\\PHD\\R with Gora Data\\R functions code from MixOmics
3_0\\Macros adapted from Mixomics_version 4.R")

### amend the location here and it will follow through all code below ###
##data location###
datloc<-"D:\\Lyns Stuff\\PHD\\3rd year plan and record of work\\GWAS modelling\\"
##output location ##
outloc<-"D:\\Lyns Stuff\\PHD\\3rd year plan and record of work\\GWAS
modelling\\independanttest\\"

#####################################################################
# inport required Y data and ensure only GWAS subjects kept        #
#  Then split the GWAS subjects into 20=test, 80%=training         #
# This is done by sorting by the larsen score & every 5th subject  #
# going into the test set                                           #
#####################################################################

yin <-paste(datloc,"mody.csv", sep="")
mody <- read.csv(yin, sep="," , header=T)
gwin <-paste(datloc,"GWAS_subj.csv", sep="")
gwas <- read.csv(gwin, sep="," , header=T)
rownames(gwas)=gwas$subjid
rownames(mody)=mody$subjid
ydata <-merge(gwas, mody, all.x = TRUE)
ydata2 <-ydata[,c(1,11)]
ydata3 <-ydata2[order(ydata2[,2]),]
ydata4=split(ydata3, rep(1:5, length = nrow(ydata3)))

testonly=ydata4[[3]]
testonly2<-testonly[order(testonly[,1]),]

train1=ydata4[[1]]
train2=ydata4[[2]]
train4=ydata4[[4]]
train5=ydata4[[5]]
trainonly<-rbind(train1,train2,train4,train5)
nrow(trainonly)
nrow(testonly)

trainonly<-trainonly[order(trainonly[,1]),]
testonly<-testonly[order(testonly[,1]),]

larsen <-as.numeric(trainonly[,2])
trainingsubj<-trainonly[,1]
larsentest<-as.numeric(testonly[,2])
testsubj<-testonly[,1]
```

```
testrows<-c(5 , 6 , 8 , 9 , 23 , 28 , 31 , 34 , 37 , 42 , 47 , 49 , 51 , 61 , 67 , 70 ,
71 , 84 , 102 , 104 , 110 , 111 , 113 , 115 , 130 , 136 , 139 , 140 , 143 , 144 , 150 ,
151 , 167 , 180 , 187 , 192 , 197 , 205 , 215 , 216 , 219 , 224 , 225 , 228 , 229 , 236
, 245 , 249 , 252 , 253 , 259 , 262 , 267 , 277 , 282 , 285 , 293 , 304 , 309 , 310 ,
314 , 316 , 318 , 323 , 327 , 335 , 336 , 342 , 347 , 348 , 349 , 357 , 358 , 373 , 376
, 377 , 379 , 383 , 387)


#####################################################################
# The following macros runs through the process of the lower level #
# modelling, each time selecting 200 variables in the PLS          #
# however the ordering of variables is continuous from 1 to max SNP#
# the n times selected though is based on being in the top 200     #
#####################################################################

rodat7 <- defmacro(chrid, n_ext=200,
                   expr={

     # inport required X dataset ###
     indata <-paste(datloc,chrid,".csv", sep="")
     modx <- read.csv(indata, sep="," , header=T)
     #then need to remove subjid from being the first variable and only keep the
training subjects
     XPRE<-modx[-testrows,-1]

#### fit the model with all variables extracted,: not limited to X vars. ####

    # set up dummy matrices to contain the output data below #
    svar<-matrix(0,ncol(XPRE),10)  # last digit here is the number of runs below
    ordSvar <-matrix(0,ncol(XPRE),10)
    nammap<-matrix("",1,ncol(XPRE))
    M=5

    ### Perform the cross validation 10 times and save the results =amend num vars to
export here####
    for (k in 1:10) {
       spls.mcv<- splscv(XPRE, larsen, ncomp=1, method="spls", mode="regression",
criterion="all",
          keepX=n_ext, validation="Mfold", M=M, max.iter=500, tol=1e-09,
          keepY=1)

### calculate the median average of the median average ranks in the 10 runs
###############
### Annoyingly ordSvar is character so export /import converts to numeric ######
      dumname<-paste(outloc,chrid,"dummy",k,".csv", sep="")
      write.csv(spls.mcv$ordSvar,file= dumname)  # outputs 1 csv per run, and all
order/ranks for the 5 folds
      temp1<-read.csv(file= dumname, sep="," , header=T)
      nammap<-temp1[,1]
      ordSvar[,k]<-apply(temp1[,2:6],1,median, na.rm=TRUE) #averages over the folds so
1-50 columns.

      ###outputs the number of times the variable is selected ###########
      svar[,k]<-as.vector(apply(spls.mcv$Svar, 1, function(a) sum(a == "TRUE")))  #add
/M*100 to get %
    }

oname<-paste(outloc,chrid,"ordSvar.csv", sep="")
write.csv(ordSvar,file= oname)
ordSvar<-read.csv(file= oname, sep="," , header=T)
ordSvar<-ordSvar[,2:11]

sname<-paste(outloc,chrid,"svar.csv", sep="")
write.csv(svar,file= sname)
svar<-read.csv(file= sname, sep="," , header=T)
svar<-svar[,2:11]

nsel<-(apply(svar,1,sum))  # number of times selected in the fold min 0, max 250 (5
folds*50 runs)
aordSvar<-apply(ordSvar,1,median, na.rm=TRUE)  #takes median rank of 50 runs,  1
obtains the row medians
```

```
bordSvar<-cbind(as.matrix(nammap),nsel,aordSvar)  #merge the names, the number times
selected & average sort order
fin_ordSvar<-bordSvar[order(aordSvar,-nsel),]  #sorts by descending n times selected &
ascending sort order

outtitle<-paste(outloc,chrid,"final",".csv",sep="")
write.csv(fin_ordSvar,file=outtitle)

outgraph<-paste(outloc,chrid,"final",".tif",sep="")
tiff(outgraph, height=1200, width=2400, res=600, units="px", pointsize=6, compression =
"lzw")
par(mar=c(5.1, 5.1, 3.1, 2.1))
plot(fin_ordSvar[,2]~fin_ordSvar[,3], xlab="Average order selected", ylab="Number of
times selected", pch=20, cex=0.3,  )
dev.off()

### Can use graph above to determine cut off for the number of variables to take
forward to the next model -top left variables##

})

rodat7("chr1a")
rodat7("chr1b")
rodat7("chr1c")
rodat7("chr2a")
rodat7("chr2b")
rodat7("chr2c")
rodat7("chr3a")
rodat7("chr3b")
rodat7("chr4a")
rodat7("chr4b")
rodat7("chr5a")
rodat7("chr5b")
rodat7("chr6a")
rodat7("chr6b")
rodat7("chr6c")
rodat7("chr7a")

rodat7("chr8a")
rodat7("chr8b")
rodat7("chr9a")
rodat7("chr9b")
rodat7("chr10a")
rodat7("chr10b")
rodat7("chr11a")
rodat7("chr11b")
rodat7("chr12a")
rodat7("chr12b")
rodat7("chr13a")
rodat7("chr13b")

rodat7("chr15")
rodat7("chr16")
rodat7("chr17")
rodat7("chr18")
rodat7("chr19")
rodat7("chr20")
rodat7("chr21")
rodat7("chr22")

rodat7("chr25", n_ext=70)


#inports new mixomics version 4 macro at 85% done 92% #####

source("D:\\Lyns Stuff\\PHD\\R with Gora Data\\R functions code from MixOmics
3_0\\Macros adapted from Mixomics_version 4b.R")
rodat7("chr7b")  # needs running at the 85% level not 92%
rodat7("chr14")  # needs running at the 85% level not 92%
rodat7("chr23", n_ext=100)  # needs running at the 85% level not 92%
```

```
##########Run all Chromosomes through above model, then choose final set of variables
based on plots ##
########## & run chosen variables through upper level model to get prediction - do we
get same variables as before? #



#############################################################################
#  This macro reads in the block of chromosome, keeps only the chosen variables  #
# above  Well those with an average order of selection less than 200            #
#############################################################################

rodat1 <- defmacro(chrid,
                  expr={

   indata <-paste(datloc,chrid,".csv", sep="")
   datx <- read.csv(indata, sep=","  , header=T)

   inkeep <-paste(outloc,chrid,"final.csv", sep="")
   keepx <- read.csv(inkeep, sep=","  , header=T)
   selvar<-keepx[which(keepx$aordSvar<200),2]
   myvars <- names(datx) %in% selvar
   chrid<- datx[,myvars]

})

rodat1(chrid="chr1a")
rodat1(chrid="chr1b")
rodat1(chrid="chr1c")
rodat1(chrid="chr2a")
rodat1(chrid="chr2b")
rodat1(chrid="chr2c")
rodat1(chrid="chr3a")
rodat1(chrid="chr3b")
rodat1(chrid="chr4a")
rodat1(chrid="chr4b")
rodat1(chrid="chr5a")
rodat1(chrid="chr5b")
rodat1(chrid="chr6a")
rodat1(chrid="chr6b")
rodat1(chrid="chr6c")
rodat1(chrid="chr7a")

rodat1(chrid="chr8a")
rodat1(chrid="chr8b")
rodat1(chrid="chr9a")
rodat1(chrid="chr9b")
rodat1(chrid="chr10a")
rodat1(chrid="chr10b")
rodat1(chrid="chr11a")
rodat1(chrid="chr11b")
rodat1(chrid="chr12a")
rodat1(chrid="chr12b")
rodat1(chrid="chr13a")
rodat1(chrid="chr13b")

rodat1(chrid="chr15")
rodat1(chrid="chr16")
rodat1(chrid="chr17")
rodat1(chrid="chr18")
rodat1(chrid="chr19")
rodat1(chrid="chr20")
rodat1(chrid="chr21")
rodat1(chrid="chr22")

rodat1(chrid="chr7b")
rodat1(chrid="chr14")
rodat1(chrid="chr23")
```

```
###now need to merge all the columns together to create 1 X dataset ###

fxvars<-
cbind(chr1a,chr1b,chr1c,chr2a,chr2b,chr2c,chr3a,chr3b,chr4a,chr4b,chr5a,chr5b,chr6a,chr
6b,chr6c,chr7a,chr7b,chr8a,chr8b,chr9a,chr9b,chr10a,chr10b,chr11a,chr11b,chr12a,chr12b,
chr13a,chr13b,chr14,chr15,chr16,chr17,chr18,chr19,chr20,chr21,chr22,chr23)
sfxvars<-unique(names(fxvars))
finxv<- fxvars[,sfxvars]
ncol(finxv)

###need to write to both out locaiton & data location as macro below always reads from
data location but will overwrite each time##
outtitle<-paste(datloc,"upperlevelmodel_final_selvars",".csv",sep="")
write.csv(finxv,file=outtitle)
outtitle<-paste(outloc,"upperlevelmodel_final_selvars",".csv",sep="")
write.csv(finxv,file=outtitle)

source("D:\\Lyns Stuff\\PHD\\R with Gora Data\\R functions code from MixOmics
3_0\\Macros adapted from Mixomics_version 4.R")

######assuming all variables above are to be kept, then run the upper level model
through extracting 1000 each time ####
rodat7("upperlevelmodel_final_selvars", n_ext=2000)



#############################################################################
#This creates CSV file of the final model, with fully ranked variables, can then #
#test the model with all... or some variables entered.  If take top 100,200,6?   #
#############################################################################

#####################################################################
##########See which model forms the best prediction 3 variables   #########
#####################################################################


   indata <-paste(outloc,"upperlevelmodel_final_selvarsfinal.csv", sep="")
   resord <- read.csv(indata, sep=","  , header=T)
   resord2<-resord[1:3,2]  # 3 is the number of vars which will be in model

   indat <-paste(outloc,"upperlevelmodel_final_selvars.csv", sep="")
   datx <- read.csv(indat, sep=","  , header=T)
   myvars <- names(datx) %in% resord2
   finaldat<- datx[-testrows,myvars]
   newdat<- datx[testrows,myvars]

##################Forms model using the final variables and the 315 subjects #
##  Then uses the newdat indp subj, to predict larsens score #######
   res <-pls(finaldat, larsen, ncomp=1, max.iter=500, tol=1e-09) #spls or pls gets same
model here#
   pred <-predict(res, newdat)


##output variables from the prediction model -beta hat are the coefficients of the
regression###
# NOTE: coefficient below (1) represents 1 component only, increase if more components
#
   vpred<-cbind(names(finaldat),round(pred$B.hat[,,1], digits=3))
   cnames<-c("xname","p_est")
   colnames(vpred)<-cnames
  vpred

       # Any predicted values <0 or >160 amended to be predicted as 0 and 160 #
       comp2<-apply(as.matrix(pred$predict[,1,1]), 2 , function(x) ifelse(x<=0,0,
ifelse(x>=160,160,x)) )
       predvals<-data.frame(cbind(comp2,larsentest))

       #plot actual value vs predicted value after X components #
outgraph<-paste(outloc,"upperlevelmodelfinal3variables_indtest",".tif",sep="")
tiff(outgraph, height=2400, width=2400, res=600, units="px", pointsize=6, compression =
"lzw")
par(mar=c(5.1, 5.1, 3.1, 2.1))
```

```
       plot (predvals$V1,predvals$larsentest, xlab="Predicted Larsen Score",
ylab="Larsen score",
            pch=4, cex=0.8, cex.axis=0.6,
           xlim=c(0,180), xaxp=c(0,180,9), ylim=c(0,180), yaxp=c(0,180,9))
       #   Get the correlation of the plot #
       corrv<-round(cor(predvals$V1,predvals$larsentest),digits=3)
       text(120,5,"r=")
       text(140,5,corrv)
       abline(1,1)

dev.off()


#############################################################################
#  This runs multiple times, using a different number of variables and plots #
# the correlation each time for the prediction on the independent set of data#
#############################################################################

  corrs <-matrix(0,100,2)

  for (k in 2:100) {

   indata <-paste(outloc,"upperlevelmodel_final_selvarsfinal.csv", sep="")
   resord <- read.csv(indata, sep=","  , header=T)
   resord2<-resord[1:k,2]

   indat <-paste(datloc,"upperlevelmodel_final_selvars.csv", sep="")
   datx <- read.csv(indat, sep=","  , header=T)
   myvars <- names(datx) %in% resord2
   finaldat<- datx[-testrows,myvars]
   newdat<- datx[testrows,myvars]

    #########Forms model using the final variables and the 315 subjects #######
    ##     Then uses the newdat indp subj, to predict larsens score #######
    res <-pls(finaldat, larsen, ncomp=1, max.iter=500, tol=1e-09)  #spls or pls gets
same model here#
    pred <-predict(res, newdat)

    ##output variables from the prediction model -beta hat are the coefficients of the
regression###
    # NOTE: coefficient below (1) represents 1 component only, increase if more
components #
    vpred<-cbind(names(finaldat),round(pred$B.hat[,,1], digits=3))
    cnames<-c("xname","p_est")
    colnames(vpred)<-cnames
    vpred

       # Any predicted values <0 or >160 amended to be predicted as 0 and 160 #
       comp2<-apply(as.matrix(pred$predict[,1,1]), 2 , function(x) ifelse(x<=0,0,
ifelse(x>=160,160,x)) )
       predvals<-data.frame(cbind(comp2,larsentest))

       corrs[k,2]<-round(cor(predvals$V1,predvals$larsentest),digits=3)
      corrs[k,1]<-k
}

corrs
corr<-corrs[-1,]

outgraph<-paste(outloc,"Correlationfordifferentvariableskeptinthemodel",".tif",sep="")
tiff(outgraph, height=2400, width=2400, res=600, units="px", pointsize=6, compression =
"lzw")
par(mar=c(5.1, 5.1, 3.1, 2.1))

plot(corr[,2]~corr[,1], pch=20, type="l", xlim=c(0,100), ylim=c(0,0.7),
xaxp=c(0,100,10), yaxp=c(0,0.7,7),
             xlab="Number of variables in the model", ylab="Correlation")
text(10,0.65,"Maximum correlation observed with 3 variables, r=0.622", adj=0)
text(10,0.62,"Disease duration, Symptom duration and Age at diagnosis", adj=0)

dev.off()
```

Appendix H: Amended 'spls' and 'valid' functions from mixOmics version 3.0 (amended version 5 for multiple Y variables and multiple components)

```
##############################################################################
#    Here are two macros which were taken from the Mixomics Version 3.0 function##
#    and adapted for specific use on the SNP data modelling the Larsen score    ##
#                                                                               ##
#  Version 1: The first updates were to output more of the datasets to be       ##
#     available for use after the macro was run                                 ##
#    For spls to spls_lyn, the changes enabled a list of the selected variables ##
#    ranked for each CV to be output into a dataset so once the model           ##
#    is selected in 8/10 folds in 50 runs, the variables can be                 ##
#    ranked in order of importance                                              ##
#                                                                               ##
#    For valid to splscv, The changes enable it to output the number of times   ##
#    each variable is selected in the M folds (Svar) and it outputs (Ycorr)     ##
#    which has the original Y values and the predicted Ys from the separate     ##
#    folds of the data                                                          ##
#                                                                               ##
# Version 2: Was written for the GWAS modelling                                 ##
#    The change was to make SNPs of 0 variance in some cross-validations just   ##
#    be excluded from that run rather than to make the analysis fall down       ##
#                                                                               ##
# Version 3: amended to include 0 variance as 0 or close to 0, as the NIPALS    ##
#    algorithm imputed 0.0001 as SNP value but this was still insufficient      ##
##   variation for the model to fit                                             ##
#                                                                               ##
# Version 4: amended code to export list of variable selection order instead    ##
#    of ranking top X variables and assigning equal last rank for those not     ##
#    selected                                                                   ##
#                                                                               ##
#                                                                               ##
# Version 5: in the case where multiple components are required for the final   ##
#    model, the order of importance is no longer simply the minimum rank of     ##
#    either of the components:selvarinord=apply(ordselvar,1,function(x) min(x)) ##
#    Instead, we want to list the order of variables contributing to each       ##
#    component.  Therefore amended macro to output a list of variable order     ##
#    for each component Svar_comp1 to ncomp and ordSvar_comp1 to ncomp          ##
#    This is for multiple Y variables as larsen score no longer Y variable it   ##
#    doesn't have its predicted values limited to within 0 & 160                ##
#                                                                               ##
# To ensure consistency, the macros were moved to this program                  ##
# to be run at the start of any analyses program thus ensuring the same code    ##
# is being used each time                                                       ##
##############################################################################


##############################################################################
#  Original macros are from the following package: Mixomics Version 3.0        #
#                                                                              #
#  LÊ CAO, K.-A., I, G. & S, D. 2009. integrOmics: an R package to unravel     #
#  relationships between two omics data sets. Bioinformatics, 25 (21), 2855-2856.#
#  NOTE: the package 'integrOmics' has been renamed 'mixOmics'.                #
#                                                                              #
# GONZÁLEZ, I., LÊ CAO, K.-A. & DÉJEAN, S. 2011. mixOmics: Omics Data Integration#
# Project. URL: http://www.math.univ-toulouse.fr/~biostat/mixOmics/.           #
##############################################################################



##############################################################################
#  MACRO 1: AMENDED spls FUNCTION and renamed spls_lyn                         #
##############################################################################

spls_lyn <-
function(X,
         Y,
         ncomp = 2,
         mode = c("regression", "canonical"),
         max.iter = 500,
```

```
          tol = 1e-06,
          keepX = rep(ncol(X), ncomp),
          keepY = rep(ncol(Y), ncomp),
          ...)
{

    #-- validation des arguments --#
    if (length(dim(X)) != 2)
        stop("'X' must be a numeric matrix.")

    X = as.matrix(X)
    Y = as.matrix(Y)

    if (!is.numeric(X) || !is.numeric(Y))
        stop("'X' and/or 'Y' must be a numeric matrix.")

    n = nrow(X)
    q = ncol(Y)

    if ((n != nrow(Y)))
        stop("unequal number of rows in 'X' and 'Y'.")

    if (is.null(ncomp) || !is.numeric(ncomp) || ncomp <= 0)
        stop("invalid number of variates, 'ncomp'.")

    nzv = nearZeroVar(X, ...)

    if (length(nzv$Position > 0)) {
        warning("Zero- or near-zero variance predictors.
  Reset predictors matrix to not near-zero variance predictors.
  See $nzv for problematic predictors.")
        X = X[, -nzv$Position]
    }

       p = ncol(X)

    ncomp = round(ncomp)
    if(ncomp > p) {
        warning("Reset maximum number of variates 'ncomp' to ncol(X) = ", p, ".")
        ncomp = p
    }

    if (length(keepX) != ncomp)
        stop("length of 'keepX' must be equal to ", ncomp, ".")

    if (length(keepY) != ncomp)
        stop("length of 'keepY' must be equal to ", ncomp, ".")

    if (any(keepX > p))
        stop("each component of 'keepX' must be lower or equal than ", p, ".")

    if (any(keepY > q))
        stop("each component of 'keepY' must be lower or equal than ", q, ".")

    mode = match.arg(mode)

    #-- initialisation des matrices --#
    X.names = dimnames(X)[[2]]
    if (is.null(X.names)) X.names = paste("X", 1:p, sep = "")

    if (dim(Y)[2] == 1) Y.names = "Y"
    else {
        Y.names = dimnames(Y)[[2]]
        if (is.null(Y.names)) Y.names = paste("Y", 1:q, sep = "")
    }

    ind.names = dimnames(X)[[1]]
    if (is.null(ind.names)) {
        ind.names = dimnames(Y)[[1]]
        rownames(X) = ind.names
    }
```

```
if (is.null(ind.names)) {
    ind.names = 1:n
    rownames(X) = rownames(Y) = ind.names
}

#-- centrer et réduire les données --#
X = scale(X, center = TRUE, scale = TRUE)
Y = scale(Y, center = TRUE, scale = TRUE)


X.temp = X
Y.temp = Y
mat.t = matrix(nrow = n, ncol = ncomp)
mat.u = matrix(nrow = n, ncol = ncomp)
mat.a = matrix(nrow = p, ncol = ncomp)
mat.b = matrix(nrow = q, ncol = ncomp)
mat.c = matrix(nrow = p, ncol = ncomp)
mat.d = matrix(nrow = q, ncol = ncomp)
   n.ones = rep(1, n)
   p.ones = rep(1, p)
   q.ones = rep(1, q)
   na.X = FALSE
na.Y = FALSE
is.na.X = is.na(X)
is.na.Y = is.na(Y)
   if (any(is.na.X)) na.X = TRUE
if (any(is.na.Y)) na.Y = TRUE
```

```
#  added by Lyn into version 1: created a blank matrix called ordselvar to
complete the ordering of selected variables by component #
```

```
ordselvar=matrix(nrow=p, ncol=ncomp)

#-- boucle sur h --#
for (h in 1:ncomp) {
    nx = p - keepX[h]
    ny = q - keepY[h]

    #-- svd de M = t(X)*Y --#
    X.aux = X.temp
    if (na.X) X.aux[is.na.X] = 0

    Y.aux = Y.temp
    if (na.Y) Y.aux[is.na.Y] = 0

    M = crossprod(X.aux, Y.aux)
    svd.M = svd(M, nu = 1, nv = 1)
    a.old = svd.M$u
    b.old = svd.M$v

    #-- latent variables --#
    if (na.X) {
        t = X.aux %*% a.old
        A = drop(a.old) %o% n.ones
        A[is.na.X] = 0
        a.norm = crossprod(A)
        t = t / diag(a.norm)
        t = t / drop(sqrt(crossprod(t)))
    }
    else {
        t = X.temp %*% a.old / drop(crossprod(a.old))
        t = t / drop(sqrt(crossprod(t)))
    }

    if (na.Y) {
        u = Y.aux %*% b.old
        B = drop(b.old) %o% n.ones
        B[is.na.Y] = 0
        b.norm = crossprod(B)
        u = u / diag(b.norm)
        u = u / drop(sqrt(crossprod(u)))
```

M=X'Y.
a.old= loadings of X (p),
b.old=loadings of Y (q)

As previously imputed missing with NIPALS algorithm this code isn't needed

t=Xp / p'p scores of X

As previously imputed missing with NIPALS algorithm this code isn't needed

```
}
else {
    u = Y.temp %*% b.old / drop(crossprod(b.old))
    u = u / drop(sqrt(crossprod(u)))
}

iter = 1

#-- boucle jusqu'à convergence de a et de b --#
repeat {
    if (na.X) a = t(X.aux) %*% u
    else a = t(X.temp) %*% u

      if (na.Y) b = t(Y.aux) %*% t
    else b = t(Y.temp) %*% t

    if (nx != 0) {
```

u=Yq/ q'q scores of Y

Calculates the loadings again from p=X'u and q=Y't

```
#Lyn added in version 3: object loadord is created to rank the loadings before
applying the 0 to those not extracted
            loadord=a

#Lyn note:abs(a[order(abs(a))][nx]) returns the value (loading coefficient) of the
#corresponding max number to be extracted, i.e. extracting 20 vars, returns the 20th
coefficient.
            a = ifelse(abs(a) > abs(a[order(abs(a))][nx]),
                (abs(a) - abs(a[order(abs(a))][nx])) * sign(a), 0)
```

Selects the top X variables to keep in the model based on the size of the p loadings

```
# Lyn added in version 1: ranks the selected variables in order of importance and
outputs them in the object ordselvar#

            ordselvar[,h]=p-(rank(abs(loadord),ties="average"))+1

        }
        a = a / drop(crossprod(u))
        a = a / drop(sqrt(crossprod(a)))

        if (ny != 0) {
            b = ifelse(abs(b) > abs(b[order(abs(b))][ny]),
                (abs(b) - abs(b[order(abs(b))][ny])) * sign(b), 0)
        }
        b = b / drop(crossprod(t))

        if (na.X) {
            t = X.aux %*% a
            A = drop(a) %o% n.ones
            A[is.na.X] = 0
            a.norm = crossprod(A)
            t = t / diag(a.norm)
            t = t / drop(sqrt(crossprod(t)))
        }
        else {
            t = X.temp %*% a / drop(crossprod(a))
            t = t / drop(sqrt(crossprod(t)))
        }

        if (na.Y) {
            u = Y.aux %*% b
            B = drop(b) %o% n.ones
            B[is.na.Y] = 0
            b.norm = crossprod(B)
            u = u / diag(b.norm)
            u = u / drop(sqrt(crossprod(u)))
        }
        else {
            u = Y.temp %*% b / drop(crossprod(b))
            u = u / drop(sqrt(crossprod(u)))
```

Selects the top Y variables to keep in the model based on the size of the q loadings

Recalculate the scores t and u based on the new loadings.

```
        }

        if (crossprod(a - a.old) < tol) break

        if (iter == max.iter) {
           warning(paste("Maximum number of iterations reached for the component",
h),
                   call. = FALSE)
            break
        }

        a.old = a
        b.old = b
        iter = iter + 1
    }

    #-- deflation des matrices --#
    if (na.X) {
        X.aux = X.temp
        X.aux[is.na.X] = 0
        c = crossprod(X.aux, t)
        T = drop(t) %o% p.ones
        T[is.na.X] = 0
        t.norm = crossprod(T)
        c = c / diag(t.norm)
    }
    else {
        c = crossprod(X.temp, t) / drop(crossprod(t))
    }

    X.temp = X.temp - t %*% t(c)

    #-- mode canonique --#
    if (mode == "canonical") {
        if (na.Y) {
            Y.aux = Y.temp
            Y.aux[is.na.Y] = 0
            e = crossprod(Y.aux, u)
            U = drop(u) %o% q.ones
            U[is.na.Y] = 0
            u.norm = crossprod(U)
            e = e / diag(u.norm)
        }
        else {
            e = crossprod(Y.temp, u) / drop(crossprod(u))
        }

        Y.temp = Y.temp - u %*% t(e)
    }

    #-- mode regression --#
    if(mode == "regression") {
        if (na.Y) {
            Y.aux = Y.temp
            Y.aux[is.na.Y] = 0
            d = crossprod(Y.aux, t)
            T = drop(t) %o% q.ones
            T[is.na.Y] = 0
            t.norm = crossprod(T)
            d = d / diag(t.norm)
        }
        else {
            d = crossprod(Y.temp, t) / drop(crossprod(t))
        }

        Y.temp = Y.temp - t %*% t(d)
    }

    mat.t[, h] = t
    mat.u[, h] = u
    mat.a[, h] = a
    mat.b[, h] = b
```

If the X loadings p are similar to the last iteration (per the Tol value) then the process is stopped.

The first component scores & loading are agreed and we deflate the X matrix using:
Xnew = X – t c' where c=X't / t't

Not doing canonical modelling this section of code not used

The first component scores & loading are agreed and we deflate the Y matrix using:
Ynew = Y – t d' where d=Y't / t't

```
        mat.c[, h] = c
        if (mode == "regression") mat.d[, h] = d

    } #-- fin boucle sur h --#

    #-- valeurs sortantes --#
    rownames(mat.a) = rownames(mat.c) = X.names
    rownames(mat.b) = Y.names
    rownames(mat.t) = rownames(mat.u) = ind.names

    dim = paste("comp", 1:ncomp)
    colnames(mat.t) = colnames(mat.u) = dim
    colnames(mat.a) = colnames(mat.b) = colnames(mat.c) = dim

    cl = match.call()
    cl[[1]] = as.name('spls')

    result = list(call = cl,
                  X = X, Y = Y, ncomp = ncomp, mode = mode,
                  keepX = keepX,
                  keepY = keepY,
                  mat.c = mat.c,
                  mat.t = mat.t,
                  variates = list(X = mat.t, Y = mat.u),
                  loadings = list(X = mat.a, Y = mat.b),
                  names = list(X = X.names, Y = Y.names, indiv = ind.names),

 # added by Lyn : version 1 if >1 component the order is the min selected in either
component.  selvarinord is output so can be used outside the function #
             selvarinord=apply(ordselvar,1,function(x) min(x))

      )
    if (length(nzv$Position > 0)) result$nzv = nzv

    class(result) = c("spls", "pls")
    return(invisible(result))
}
```

```
#############################################################################
#  MACRO 2: AMENDED valid FUNCTION renamed splscv                           #
#############################################################################
```

```
splscv <-
function(X,
         Y,
         ncomp = min(6, ncol(X)),
         method = "spls",
         mode = c("regression"),
         criterion = c("all"),
         keepX = NULL, keepY = NULL,
         validation = c("Mfold"),
         M = M,
         max.iter = 500,
         tol = 1e-06, ...)
{

    method = match.arg(method)

    #-------------------- sPLS --------------------#
    if (any(c("spls") == method)) {

        #-- validation des arguments --#
        #-- do warning for mode + other warnings --#
        if (length(dim(X)) != 2)
            stop("'X' must be a numeric matrix.")

        mode = match.arg(mode)
```

```
        validation = match.arg(validation)

        X = as.matrix(X)
        Y = as.matrix(Y)

        n = nrow(X)
        q = ncol(Y)
        res = list()

        if (!is.numeric(X) || !is.numeric(Y))
            stop("'X' and/or 'Y' must be a numeric matrix.")

        if ((n != nrow(Y)))
            stop("unequal number of rows in 'X' and 'Y'.")

        if (any(is.na(X)) || any(is.na(Y)))
            stop("Missing data in 'X' and/or 'Y'. Use 'nipals' for dealing with NAs.")
```

```
# Added by Lyn: Version 2
# REMOVED the following check as variables with insufficient variation will be removed
on
# a variable by variable basis according to the cross folds used. See below:
#  We do not want them excluded from all analyses.

#             nzv = nearZeroVar(X, ...)
#        if (length(nzv$Position > 0)) {
#            warning("Zero- or near-zero variance predictors.
#  Reset predictors matrix to not near-zero variance predictors.
#  See $nzv for problematic predictors.")
#            X = X[, -nzv$Position]
#                res$nzv = nzv
#        }
```

```
           p = ncol(X)

        if (is.null(ncomp) || !is.numeric(ncomp) || ncomp <= 0 || ncomp > p)
            stop("Invalid number of components, 'ncomp'.")
        ncomp = round(ncomp)

        if (method == "spls") {
            if(is.null(keepX)) keepX = rep(ncol(X), ncomp)
            if(is.null(keepY)) keepY = rep(ncol(Y), ncomp)

            if (length(keepX) != ncomp)
                stop("length of 'keepX' must be equal to ", ncomp, ".")

            if (length(keepY) != ncomp)
                stop("length of 'keepY' must be equal to ", ncomp, ".")

            if (any(keepX > p))
                stop("each component of 'keepX' must be lower or equal than ", p, ".")

            if (any(keepY > q))
                stop("each component of 'keepY' must be lower or equal than ", q, ".")

        }

        #-- M fold  validation --#
        ##- define the folds
        if (validation == "Mfold") {
            if (is.null(M) | !is.numeric(M) | M < 2 | M > n)
                stop("Invalid number of folds, 'M'.")
            M = round(M)
            fold = split(sample(1:n), rep(1:M, length = n))
        }

        #-- compute MSEP and/or R2 --#
        if (any(criterion %in% c("all", "MSEP", "R2"))) {
            press.mat = Ypred = array(0, c(n, q, ncomp))
            MSEP = R2 = matrix(0, nrow = q, ncol = ncomp)
```

```
#  Lyn added 3 rows (for svar, Ycorr and ordSvar) in version 1 to create a blank
matrices # for filling in, however these were amended in version 5 to allow for
multiple components
# This is the number Y variables (q) times 2 as got actual value & predicted value in
the # matrix.  Svar matrix is created containing the order of variables for all
components
```

```
             Ycorr=matrix(0, n, q*2)
             assign(paste("ordSvar", sep = ""), matrix("0",p,(ncomp*M)+1)  )
             assign(paste("Svar", sep = ""), matrix("0", ncol(X), (ncomp*M)+1) )

            for (i in 1:M) {
                omit = fold[[i]]
                X.train = X[-omit, ]
                Y.train = Y[-omit, ]
                X.test = matrix(X[omit, ], nrow = length(omit))
                  Y.test = matrix(Y[omit, ], nrow = length(omit))
```

```
## added by lyn: version 2 because there are SNPs which when folded have no variation
##  all 0's, then don't want process to fall down, so just exclude that variable
##  from that fold.  If a variable is predictive, then it will come up in all 50 runs
## but not in all folds as it has too low MAF with the sample size.   Lyn has
## programmed to include variables with at least 92% variation in the training set
## so minor alleles with <5% MAF won't be included
```

```
# Lyn added for version 3:  Minor code change from version 2:
# From sum(X>==0) to sum(x>=0 & x<=0.5).  This allows for 0'S or very small close to
zero # numbers derived from the NIPALS algorithm missing data imputation
# Otherwise macro still falls down with insufficient variation
```

```
        MAF <-apply(X.train[,1:ncol(X.train)], 2, function(x) sum(x>=0 & x<=0.5)) /
nrow(X.train[,1:ncol(X.train)]) * 100
            MAF2<-MAF>92  # put TRUE to those with >92% 0s.

###remove any with >92% 0s from the training & test X data ###
            X.train<-X.train[,!MAF2]
            X.test<-X.test[,!MAF2]


                X.train = scale(X.train, center = TRUE, scale = FALSE)
                xmns = attr(X.train, "scaled:center")

                Y.train = scale(Y.train, center = TRUE, scale = FALSE)
                ymns = attr(Y.train, "scaled:center")

                X.test = scale(X.test, center = xmns, scale = FALSE)

                #-- spls --#
                object = spls_lyn(X = X.train, Y = Y.train, ncomp = ncomp,
                                  mode = mode, max.iter = max.iter, tol = tol,
                                  keepX = keepX, keepY = keepY)
```

```
####################Lyn amended the code here : in version 1###########
##    It will now export the predicted scores, the B coefficients     ##
#     and will list the variables selected                            ##
##    For each of the M models, for h components                      ##
##    It will also export the ordering of the selected variables      ##
#########################################################################
```

```
          predpar = predict(object,X.test)
          Y.hat=predpar$predict

          s.var=apply(abs(object$loadings$X),1,sum)>0
```

This tells us what is in the model and which variables are out

```
#  Lyn added version 2: As some variables excluded due to insufficient variation in
some #folds, the following code was added so that a complete list of variables in
created.
# Those not included as set to 0.
```

```
                MAF3<-cbind(MAF2,names(MAF2))
                colnames(MAF3)<-c("EIMOD","SNP")
```

```
# Added into version 5, instead of taking sum across all loadings for all coefficients
# keep separate and count number of times in /out of model for each component
# svar shows which variable is in model and which out for each comp (s)
# for 1 comp will be in columns 2 to m+1, 2 comp will be in m+2 to 2m+1 columns. etc.
```

```
          for (s in 1:ncomp) {
                s.var<-apply(as.matrix(abs(object$loadings$X[,s] )),1,sum)>0
                s.var2<-cbind(s.var,names(s.var))
                colnames(s.var2)<-c("CBMOD","SNP")

                s.var3 <- merge(s.var2,MAF3, by="SNP", all=TRUE)
                s.var3 [is.na(s.var3)] <- FALSE
                s.var5<-as.matrix(s.var3[,2])
                rownames(s.var5)<-s.var3[,1]
                s.var6<- s.var5[order(row.names(s.var5)),]

                Svar[,1]<-as.matrix(names(s.var6))
                Svar[,(((s*M)+i)-M)+1]<-as.matrix(s.var6)

           }
```

```
# Lyn amended here in Version 2: need to label selvarinord with var names.
# Merges on the variables which were not fitted above and get a complete list
# which all folds can be merged by.

# Amended again in version 5, so that a matrix is created for each component containing
the order of variables
```

```
            for (s in 1:ncomp) {
                revar<-names(MAF2[!MAF2])
                selvarinord2<-cbind(object$ordselvar[,s],revar)   #selvarinord =min
order across comps, ordselvar=1 for each comp
                colnames(selvarinord2)<-c("CHORD","SNP")
                selvarinord3<-merge(selvarinord2, MAF3, by="SNP", all=TRUE)
                selvarinord4<-as.matrix(selvarinord3[,2])
                rownames(selvarinord4)<-selvarinord3[,1]
                selvarinord5<- selvarinord4[order(row.names(selvarinord4)),]

                ordSvar[,1]<-as.matrix(names(selvarinord5))
                ordSvar[,(((s*M)+i)-M)+1]<-as.matrix(selvarinord5)
           }
```

```
               for (h in 1:ncomp) {
                      Y.mat = matrix(Y.hat[, , h], nrow = dim(Y.hat)[1], ncol=
dim(Y.hat)[2])
                  Y.hat[, , h] = sweep(Y.mat, 2, ymns, FUN = "+")
```

```
#  Lyn amended in version 5 as multiple Y variables, so do not want to restrict the
prediction to be 0-160 as different variables in different units.

# Y.hat[, , h]<-apply(as.matrix(Y.hat[, , h]), 2 , function(x) ifelse(x<=0,0,
ifelse(x>=160,160,x)) )
```

```
                press.mat[omit, , h] = (Y.test - Y.hat[, , h])^2
                Ypred[omit, , h] = Y.hat[, , h]
               }
```

```
# Added by Lyn in version 1 but amended again in version 5, to calculate the
correlation
# between the predicted Y vars (using all components) and the actual Y values
# When more than 1 component, need to sum the predicted Y.hats to make the complete
model
```

```
               gpredval<-apply(simplify2array(Y.hat), c(1,2), sum)
               Ycorr[omit,]<-cbind(as.matrix(gpredval), as.matrix(Y.test))
               prefixc<-rep("p_",q)   #prefix predicted values with a p_
               lstcolname<-colnames(Y.hat)
               lstcolname2<-paste(prefixc,lstcolname,sep="")
```

```
            colnames(Ycorr)<-c(lstcolname2,colnames(Y.hat))

        } #end i

        for (h in 1:ncomp) {
            MSEP[, h] = apply(as.matrix(press.mat[, , h]), 2, mean, na.rm = TRUE)
            R2[, h] = diag(cor(Y, Ypred[, , h], use = "pairwise"))
        }

        colnames(MSEP) = colnames(R2) = paste('ncomp', c(1:ncomp), sep = " ")
        rownames(MSEP) = rownames(R2) = colnames(Y)

        if (q == 1) rownames(MSEP) = rownames(R2) = ""

        #-- valeurs sortantes --#
        if (any(criterion %in% c("all", "MSEP"))) res$MSEP = MSEP
         if (any(criterion %in% c("all", "R2"))) res$R2 = R2

         res$Ycorr=Ycorr # line added by lyn in version 1.
  ##########following added into version 5 to output the required variables for all
components ####
         res$ordSvar<-ordSvar
         res$Svar<-Svar

       }

# calculation of Q2 removed from Version 5 as we only use the R2 under CV.
       #-- compute Q2 --#
    }


    method = paste(method, "mthd", sep = ".")
    class(res) = c("valid", method)
    return(invisible(res))
}
```

Appendix I: Code to produce GWAS analysis for multiple Y variables using 80% of data for training and independently 20% to test.

```
#-----------------------------------------------------------------------#
# 2013: Program fits SPLS analysis for the 3rd group of & variables      #
# All subjects MODELLING 80% and testing it on 20%                       #
# This program performs the cross validation  with imputed data          #
# As smaller dataset, PLS in 1 model & top X variables explored under CV #
#                                                                        #
#-----------------------------------------------------------------------#
require(mixOmics)
require(gtools)

rm(list=ls())

#### read in the macros I've adapted from mixomics ##############
source("D:\\Lyns Stuff\\PHD\\R with Gora Data\\R functions code from MixOmics
3_0\\Macros adapted from Mixomics_version 5.R")

### amend the location here and it will follow through all code below ###
##data location###
datloc<-"D:\\Lyns Stuff\\PHD\\3rd year plan and record of work\\Multiple Y modelling\\"
##output location ##
outloc<-"D:\\Lyns Stuff\\PHD\\3rd year plan and record of work\\Multiple Y modelling\\"


#####################################################################
#####################################################################
##### "DATA INPUT"                                             #####
##### PVAS, MHAQ, RASEV, ESR, CRP, Any erosions, Larsen , hand #####
#####  & foot counts + DAS28 variables to analyse              #####
#####################################################################
#####################################################################

yvars<-paste(datloc,"nipals_dasyvars_training.csv", sep="")
ydata <- read.csv(yvars, sep=","  , header=T)

xvars<-paste(datloc,"nipals_dasxvars_training.csv", sep="")
xdata <- read.csv(xvars, sep=","  , header=T)
nrow(xdata)
nrow(ydata)

#names(xdata)
xdata2<-xdata[,c(-1,-4)]
ncol(xdata2)

#names(ydata)
ydata2<-ydata[,c(2,3,4,9,10,11,12,13,14,15)]
ncol(ydata2)

#####################################################################
# Run a model with 5, 10 and 100 variables to see if 1 or 2        #
# components are needed (or more!)                                  #
#####################################################################

## Set number of components to 2
ncomp <- 2

## Total number of selected genes on all ncomp dimensions
kpX <-c(3,4,5,10,100)
R2_1<-R2_2 <-matrix(NA,length(kpX),ncol(ydata2))

for (i in 1:length(kpX)) {
    error <- splscv(xdata2, ydata2, ncomp = ncomp, keepX = rep(kpX[i],ncomp),
                        method = "spls", mode="regression", criterion="all",
                        validation = "Mfold", M = 7, max.iter=500, tol=1e-09,
                        keepY=rep(ncol(ydata2),ncomp)  )

   R2_1[i,]   <-as.vector(error$R2[,1]) # = 1st comp R2 calculated under CV
   R2_2[i,]   <-as.vector(error$R2[,2]) # = 2nd comp  R2 calculated under CV
```

```
}

###Want to compare the R2 for each number of variables extracted to see if the 2nd
component is any benefit ###
## Columns are the Y variables (as we get an R2 for each variable ####
R2<-R2_2-R2_1
anyimp<-any(R2>=0.0975)
anyimp  ## If this is false then none of the Y variables (using the 3 different number
of X variables) require 2 components.
colnames(R2)<-names(ydata2)
R2

#####################################################################
# The following macros runs through the                             #
# modelling, each time selecting 100 variables in the PLS           #
# however the ordering of variables is continuous from 1 to max SNP#
# the n times selected though is based on being in the top 100      #
#####################################################################

## Set number of components based on decision made above, 7 fold & 10 runs
ncomp <- 1
M=7
nruns=10

    # set up dummy matrices to contain the output data below #
    svar1runs<-matrix(0,ncol(xdata2),nruns)
    mord1<-matrix(0,ncol(xdata2),M)
    mord1runs <-matrix(0,ncol(xdata2),nruns)

    ### Perform the cross validation 10 times and save the results =amend num vars to
export here####
    for (k in 1:nruns) {
       spls.mcv<- splscv(xdata2, ydata2, ncomp=ncomp, method="spls", mode="regression",
criterion="all",
          keepX=rep(100,ncomp), validation="Mfold", M=M, max.iter=500, tol=1e-09,
          keepY=rep(ncol(ydata2),ncomp))

      ### calculate the median average of the median average ranks in the 10 runs for 2
components ##############
      mord1[,1:M]<-as.numeric(spls.mcv$ordSvar[,2: (M+1)])
      rownames(mord1)<-spls.mcv$ordSvar[,1]   #  Component 1, ranks for each fold.
      mord1runs[,k]<-apply(mord1,1,median, na.rm=TRUE)

      ###outputs the number of times the variable is selected ###########
      svar1runs[,k]<-as.vector(apply(spls.mcv$Svar[,2: (M+1)], 1, function(a) sum(a ==
"TRUE")))  #add /M*100 to get %
    #  svar2runs[,k]<-as.vector(apply(spls.mcv$Svar[,(M+2):(M+M+1)], 1, function(a)
sum(a == "TRUE")))
    }

nsel1<-(apply(svar1runs,1,sum))  # number of times selected in the folds & runs for
comp 1
ord1<-apply(mord1runs,1,median, na.rm=TRUE)  #takes median rank of runs,

sumstats<-cbind(as.matrix(spls.mcv$ordSvar[,1]),nsel1,ord1)  #merge the names, the
number times selected & average sort order - 1 comp

finaldset<-sumstats[order(ord1,-nsel1),]  #sorts by comp 1 average rank and descending
number of times selected - 1 comp

outtitle<-paste(outloc,"finalvarsortorderallplusdas_1comp",".csv",sep="")
write.csv(finaldset,file=outtitle)



#####################################################################
##### "TEST DATA INPUT"  ############################################
#####################################################################


ytst<-paste(datloc,"nipals_dasyvars_test.csv", sep="")
```

```
ytest <- read.csv(ytst, sep="," , header=T)

xtst<-paste(datloc,"nipals_dasxvars_test.csv", sep="")
xtest <- read.csv(xtst, sep="," , header=T)
nrow(xtest)
nrow(ytest)

#names(xtest)
xtest2<-xtest[,c(-1,-4)]
ncol(xtest2)

#names(ytest)
ytest2<-ytest[,c(2,3,4,9,10,11,12,13,14,15)]
ncol(ytest2)

#############################################################################
#  This runs multiple times, using a different number of variables and plots #
# the correlation each time for the prediction on the independent set of data#
#############################################################################

  corrs <-matrix(0,99,(ncol(ytest2)+2))

  for (k in 2:100) {

   indata <-paste(outloc,"finalvarsortorderallplusdas_1comp.csv", sep="")
   resord <- read.csv(indata, sep="," , header=T)
   resord2<-resord[1:k,2]

   myvars <- names(xdata2) %in% resord2
   finaldat<- xdata2[,myvars]
   xtestdat<- xtest2[,myvars]

    ##  Forms model using the final variables and 730 subjects #######
    ##  Then uses the newdat independent subj, to predict larsens score #######
    res <-pls(finaldat, ydata2, ncomp=1, max.iter=500, tol=1e-09)  #spls or pls gets
same model here#
    pred <-predict(res, xtestdat)

     gpredval<-apply(simplify2array(pred$predict), c(1,2), sum)  ## this averages over
2 or more components to get full prediction value
     rval<-cbind(as.matrix(gpredval), as.matrix(ytest2))
     prefixc<-rep("p_",ncol(pred$predict))   #prefix predicted values with a p_
     lstcolname<-colnames(pred$predict)
     lstcolname2<-paste(prefixc,lstcolname,sep="")
     colnames(rval)<-c(lstcolname2,colnames(ytest2))

     for (j in 1:ncol(ytest2)) {
     corrs[k-1,j+2]<-round(cor(rval[,j],rval[,j+ncol(ytest2)]),digits=3)  # calculates
average correlation across all Y variables.
     corrs[k-1,1]<-k
     colnames(corrs)<-c("Nvars","Average corr",colnames(ytest2))
     }
}

corrs[,2]<-round(rowMeans(abs(corrs[,3:(ncol(ytest2)+2)]), dims=1,
na.rm=FALSE),digits=3)

outgraph<-paste(outloc,"Correlation_allplusdas",".tif",sep="")
tiff(outgraph, height=2400, width=2400, res=600, units="px", pointsize=6, compression =
"lzw")
par(mar=c(5.1, 5.1, 3.1, 2.1))
par(mfrow=c(4,3))

names<-c("None","Average","DAS 28", "PVAS", "ESR", "MHAQ", "Foot erosions", "Hand
erosions", "Larsen score", "Any erosions", "RA severity","CRP")

for (i in 2:12) plot(corrs[,i]~corrs[,1], col="black", pch=20, type="l", xlim=c(0,100),
xaxp=c(0,100,10),
             xlab="Number of variables in the model", ylab="Correlation",
main=paste(names[i]) )
dev.off()
```

```
#### output the number of vars corresponding to the maximum correlation for each Y
variable ###
sumdset<-matrix(0,11,2)
for (i in 2:12) {
      sumdset[(i-1),2]=max(corrs[,i])
      sumdset[(i-1),1]=corrs[corrs[,i]==max(corrs[,i]),1]
}
sumdset


   indata <-paste(outloc,"finalvarsortordersjctjcdas_1comp.csv", sep="")
   resord <- read.csv(indata, sep="," , header=T)
   resord2<-resord[1:12,2]
resord2


#####################################################################
#####################################################################
#       add in a second component & refit using the 11 variables   #
#     selected from the 1 component model                           #
#####################################################################
#####################################################################


#####################################################################
# The following macros runs through the                             #
# modelling, each time selecting 100 variables in the PLS           #
# however the ordering of variables is continuous from 1 to max SNP#
# the n times selected though is based on being in the top 100      #
#####################################################################

## Set number of components based on decision made above, 7 fold & 10 runs
ncomp <- 2
M=7
nruns=10

    # set up dummy matrices to contain the output data below #
    svar1runs<-matrix(0,ncol(xdata2),nruns)
    svar2runs<-matrix(0,ncol(xdata2),nruns)  # a dataset for each component
    mord1<-matrix(0,ncol(xdata2),M)
    mord2<-matrix(0,ncol(xdata2),M)
    mord1runs <-matrix(0,ncol(xdata2),nruns)
    mord2runs <-matrix(0,ncol(xdata2),nruns)

    ### Perform the cross validation 10 times and save the results =amend num vars to
export here####
    for (k in 1:nruns) {
       spls.mcv<- splscv(xdata2, ydata2, ncomp=ncomp, method="spls", mode="regression",
criterion="all",
          keepX=c(11,100), validation="Mfold", M=M, max.iter=500, tol=1e-09,
          keepY=rep(ncol(ydata2),ncomp))

      ### calculate the median average of the median average ranks in the 10 runs for 2
components ##############
      mord1[,1:M]<-as.numeric(spls.mcv$ordSvar[,2: (M+1)])
      rownames(mord1)<-spls.mcv$ordSvar[,1]   #  Component 1, ranks for each fold.
      mord2[,1:M]<-as.numeric(spls.mcv$ordSvar[,(M+2):(M+M+1)])
      rownames(mord2)<-spls.mcv$ordSvar[,1]   #  Component 2, ranks for each fold.
      mord1runs[,k]<-apply(mord1,1,median, na.rm=TRUE)
      mord2runs[,k]<-apply(mord2,1,median, na.rm=TRUE)

      ###outputs the number of times the variable is selected ###########
      svar1runs[,k]<-as.vector(apply(spls.mcv$Svar[,2: (M+1)], 1, function(a) sum(a ==
"TRUE")))  #add /M*100 to get %
      svar2runs[,k]<-as.vector(apply(spls.mcv$Svar[,(M+2):(M+M+1)], 1, function(a)
sum(a == "TRUE")))
    }

nsel1<-(apply(svar1runs,1,sum))  # number of times selected in the folds & runs for
comp 1
```

```
nsel2<-(apply(svar2runs,1,sum))  # number of times selected in the folds & runs for
comp 2
ord1<-apply(mord1runs,1,median, na.rm=TRUE)  #takes median rank of runs,
ord2<-apply(mord2runs,1,median, na.rm=TRUE)  #takes median rank of runs,

sumstats<-cbind(as.matrix(spls.mcv$ordSvar[,1]),nsel1,nsel2,ord1,ord2)  #merge the
names, the number times selected & average sort order - 2 comp

finaldset<-sumstats[order(ord1,-nsel1,ord2,-nsel2),]  #sorts by comp 1 average rank and
descending number of times selected - 2 comp

outtitle<-paste(outloc,"finalvarsortorderallplusdas_2comp",".csv",sep="")
write.csv(finaldset,file=outtitle)


#####################################################################
##### "TEST DATA INPUT"  ############################################
#####################################################################


ytst<-paste(datloc,"nipals_dasyvars_test.csv", sep="")
ytest <- read.csv(ytst, sep=","  , header=T)

xtst<-paste(datloc,"nipals_dasxvars_test.csv", sep="")
xtest <- read.csv(xtst, sep=","  , header=T)
nrow(xtest)
nrow(ytest)

#names(xtest)
xtest2<-xtest[,c(-1,-4)]
ncol(xtest2)

#names(ytest)
ytest2<-ytest[,c(2,3,4,9,10,11,12,13,14,15)]
ncol(ytest2)

#############################################################################
#  This runs multiple times, using a different number of variables and plots #
# the correlation each time for the prediction on the independant set of data#
#############################################################################

  corrs <-matrix(0,99,(ncol(ytest2)+2))

  for (k in 2:100) {

   indata <-paste(outloc,"finalvarsortorderallplusdas_2comp.csv", sep="")
   resord <- read.csv(indata, sep=","  , header=T)
#####need to keep the top 11 from the comp1 ###
   resord1<-as.vector(resord[1:11,2] )
#####Then sort by the 2nd comp vars and extract the top 2 to 100- removing any
duplicates across the 2 components###
   resord1b<-resord[order(resord$ord2,-resord$nsel2),]
   resord2<-as.vector(resord1b[1:k,2] )
   resord3<-unique(c(resord1,resord2))
   newk<-length(resord3)

   myvars <- names(xdata2) %in% resord3
   finaldat<- xdata2[,myvars]
   xtestdat<- xtest2[,myvars]

    ##  Forms model using the final variables and 730 subjects #######
    ##  Then uses the newdat independant subj, to predict larsens score #######
    res <-pls(finaldat, ydata2, ncomp=2, max.iter=500, tol=1e-09)  #spls or pls gets
same model here#
    pred <-predict(res, xtestdat)

     gpredval<-apply(simplify2array(pred$predict), c(1,2), sum)  ## this averages over
2 or more components to get full prediction value
     rval<-cbind(as.matrix(gpredval), as.matrix(ytest2))
     prefixc<-rep("p_",ncol(pred$predict))   #prefix predicted values with a p_
     lstcolname<-colnames(pred$predict)
     lstcolname2<-paste(prefixc,lstcolname,sep="")
```

```
     colnames(rval)<-c(lstcolname2,colnames(ytest2))

     for (j in 1:ncol(ytest2)) {
     corrs[k-1,j+2]<-round(cor(rval[,j],rval[,j+ncol(ytest2)]),digits=3)  # calculates
average correlation across all Y variables.
     corrs[k-1,1]<-newk  # as adding in extra comp 2 vars which might be duplicates to
the 1st, n vars = 11+ number in 2nd comp removing dups.
     colnames(corrs)<-c("Nvars","Average corr",colnames(ytest2))
     }
}

corrs[,2]<-round(rowMeans(abs(corrs[,3:(ncol(ytest2)+2)]), dims=1,
na.rm=FALSE),digits=3)

outgraph<-paste(outloc,"Correlation_allplusdas_2comp",".tif",sep="")
tiff(outgraph, height=2400, width=2400, res=600, units="px", pointsize=6, compression =
"lzw")
par(mar=c(5.1, 5.1, 3.1, 2.1))
par(mfrow=c(4,3))

names<-c("None","Average","DAS 28", "PVAS", "ESR", "MHAQ", "Foot erosions", "Hand
erosions", "Larsen score", "Any erosions", "RA severity","CRP")

for (i in 2:12) plot(corrs[,i]~corrs[,1], col="black", pch=20, type="l", xlim=c(0,100),
xaxp=c(0,100,10),
             xlab="Number of variables in the model", ylab="Correlation",
main=paste(names[i]) )
dev.off()


#### output the number of vars corresponding to the maximum correlation for each Y
variable ###
sumdset<-matrix(0,11,2)
for (i in 2:12) {
      sumdset[(i-1),2]=max(corrs[,i])
      temp<-corrs[corrs[,i]==max(corrs[,i]),1]
      sumdset[(i-1),1]=min(temp)
}
sumdset



############################################################
# Use program below to produce scatter plots of final model#
############################################################


   indata <-paste(outloc,"finalvarsortorderallplusdas_2comp.csv", sep="")
   resord <- read.csv(indata, sep="," , header=T)
#####need to keep the top 11 from the comp1 ###
   resord1<-as.vector(resord[1:11,2] )
#####Then sort by the 2nd comp vars and extract the top 2 to 100- removing any
duplicates across the 2 components###
   resord1b<-resord[order(resord$ord2,-resord$nsel2),]
   resord2<-as.vector(resord1b[1:7,2] )  #### enter here the number of additional
vairables we decided to add in for 2nd comp. N=7.
   resord3<-unique(c(resord1,resord2))
   newk<-length(resord3)

   myvars <- names(xdata2) %in% resord3
   finaldat<- xdata2[,myvars]
   xtestdat<- xtest2[,myvars]

    ##  Forms model using the final variables and 730 subjects #######
    ##  Then uses the newdat independant subj, to predict larsens score #######
    res <-pls(finaldat, ydata2, ncomp=2, max.iter=500, tol=1e-09)  #spls or pls gets
same model here#
    pred <-predict(res, xtestdat)

     gpredval<-apply(simplify2array(pred$predict), c(1,2), sum)  ## this averages over
2 or more components to get full prediction value
     rval<-cbind(as.matrix(gpredval), as.matrix(ytest2))
```

```
     prefixc<-rep("p_",ncol(pred$predict))   #prefix predicted values with a p_
     lstcolname<-colnames(pred$predict)
     lstcolname2<-paste(prefixc,lstcolname,sep="")
     colnames(rval)<-c(lstcolname2,colnames(ytest2))

outtitle<-paste(outloc,"predvaluesgroup3_2comp",".csv",sep="")
write.csv(rval,file=outtitle)


###########plot all the correlations from the final models #######

outgraph<-paste(outloc,"RawCorrelation_allplusdas_2comp",".tif",sep="")
tiff(outgraph, height=2400, width=2400, res=600, units="px", pointsize=6, compression =
"lzw")
par(mar=c(5.1, 5.1, 3.1, 2.1))
par(mfrow=c(4,3))

names<-c("DAS 28", "PVAS", "ESR", "MHAQ", "Foot erosions", "Hand erosions", "Larsen
score", "Any erosions", "RA severity","CRP")

for (i in 1:10) {
   plot(rval[,i]~rval[,i+10], col="black", pch=20, xlab="Actual", ylab="Predicted",
main=paste(names[i]) )
   cval=cor(as.vector(rval[,i]),as.vector(rval[,i+10]))
   mtext(side=1, line=2.8, text=paste("r=",round(cval,digits=3),sep=""), adj=1,
cex=0.6)
   reg1 <- lm(rval[,i]~rval[,i+10])
   abline(reg1)
}

dev.off()
```