Burkitt lymphoma classification and *MYC*-associated non-Burkitt lymphoma investigation based on gene expression

Chulin Sha

Submitted in accordance with the requirements for the degree of

Doctor of Philosophy

The University of Leeds
Faculty of Biology
School of Molecular and Cellular Biology

April 2015

Intellectual Property and Publication Statements

1st Authored, used in this thesis,

Transferring genomics to the clinic: distinguishing Burkitt and diffuse large B cell lymphomas. Accepted for publication in Genome Medicine.

The candidate confirms that the work submitted is her own and that appropriate credit has been given where reference has been made to the work of others.

Acknowledgements

I would like to give my sincerest gratitude to my supervisor, David Westhead, for introducing me to Bioinformatics and Lymphoma research area and offering me guidance and care whenever I get confused or frustrated, particularly for his patience and encouragement when the project is going slow and for always considering my benefit where I am short of experience. Being honest, I feel so lucky to pursue my study here that I couldn't possibly have a better supervisor.

I also want to thank to my collaborators from Leeds St. James Hospital Sharon, Andrew and Reuben for providing the data, advice and help on how to be involved in a project and cooperate with others.

My dear lab mates, past and present, thanks for their companionship and providing a fantastic research atmosphere. Especially a big thank to Matt and Vijay, who help me along my whole Phd study, offering valuable ideas, encouragement and friendship. I hope I can grow to a good bioinformatician like them in the near future.

Thank to my close friends for sharing my feelings so that I don't feel so lonely and isolated. Thank to my parents and my boyfriend for their unconditional support and endless love, which make me believe I can accomplish my objectives.

At last, thank to China Scholarship Council and University of Leeds for the financial support, and to anyone who has helped me in any way during my Phd study period.

Abstract

Burkitt lymphoma and diffuse large B-cell lymphoma are two closely related types of lymphoma that are managed differently in clinical practice and the accurate diagnosis is a key point in treatment decisions. However based on current criteria combined with morphological, immunophenotypic and genetic characteristics, a significant number of cases exhibit overlapping features where diagnosis and treatment decisions are difficult to make. Especially, the prognosis have been reported significantly unfavourable in a subset of cases that are initially diagnosed as diffuse large B-cell lymphoma but bear *MYC* gene translocation, which is a defining feature of Burkitt lymphoma however can also be found in other lymphomas. Despite the adverse effect of MYC in aggressive lymphomas other than Burkitt lymphoma, the underlying mechanism and effective treatment is still unclear.

Recent technological advances have made it possible to simultaneously investigate an enormous number of bio-molecules, and the scientific fields associated with measuring molecular data in such a high-throughput way are usually called “omics”. For example, genomics assesses thousands of DNA sequences and transcriptomics assays large numbers of transcripts in a single experiment. These techniques together with the rapidly emerging analytical methods in bioinformatics have introduced cancer research into a new era. The growing amount of omics data have significantly influenced the understanding of lymphomas and hold great promise in classifying subtypes, predicting treatment responses that will eventually lead to personalized therapy.

Here in this study, we investigate the discrimination of Burkitt lymphoma and diffuse large B-cell lymphoma based on DNA microarray gene expression data, which has contributed most in molecular classification of lymphoma subtypes in the last decade. On the basis of two previous research level gene expression profiling classifiers, we developed a robust classifier that works effectively on different platforms and formalin fixed paraffin-embedded samples commonly used in routine clinic. The validation of the classifier on

the samples from clinical patients achieves a high agreement with diagnosis made in a central haematopathology laboratory, and leads to a potential outcome indication in the patients presenting intermediate features. In addition, we explore the role of *MYC* in the above lymphomas. Our investigation emphasizes the inferior impact of high level *MYC* mRNA expression on patients' outcome, and the functional analysis of *MYC* high expression associated genes show significantly enriched molecular mechanisms of proliferation and metabolic process. Moreover, the gene *PRMT5* is found to be highly correlated with *MYC* expression which opens a possible therapeutic target for the treatment.

Table of Contents

Intellectual Property and Publication Statements ..... **ii**
Acknowledgements ..... **iii**
Abstract ..... **iv**
List of Abbreviations ..... **x**
List of Tables ..... **xiii**
List of Figures ..... **xv**
Chapter 1 Introduction **1**
1.1 Burkitt lymphoma and diffuse large B-cell lymphoma 2
1.1.1. Burkitt lymphoma 2
1.1.2. Diffuse large B-cell lymphoma 4
1.1.3. Cases with features intermediate between BL and DLBCL 5
1.2 *MYC* in BCLU 7
1.2.1. *MYC* translocation 8
1.2.2. *MYC* prognostic implication 9
1.2.3. *MYC* potential mechanism in B-cell lymphoma 10
1.3 High-throughput data in cancer research 13
1.3.1. New methods and techniques 14
1.3.2. Omics-based tests translation from research to clinic16
1.4 Microarray gene expression profiling 17
1.4.1. DNA microarray technology 18
1.4.2. General analysis of DNA microarray data 20
1.4.3. Reality in DNA microarray GEP analysis 22
1.5 Research overview 23
1.5.1. Study design 23
1.5.2. Developing environment and tools 26
1.5.3. Data collection and collaboration 27
Chapter 2 Methods **29**
2.1 Low level analysis 29
2.1.1. Quality check 29
2.1.2. Preprocessing 30
2.1.3. Cross-platform normalization 31

2.2 Feature selection 32
2.2.1. Introduction 33
2.2.2. SAM 34
2.2.3. Smyth moderated t-statistic 35
2.3 Classification methods 36
2.3.1. Introduction 37
2.3.2. Support vector machines 38
2.3.3. Evaluation of classifier 40
2.4 Survival analysis 41
2.4.1Introduction 42
2.4.2Kaplan-Meier survival estimate 43
2.4.3The cox proportional hazard model 44
2.5 Mechanism analysis 45
2.5.1. Gene set enrichment analysis tool 45
2.5.2. DAVID functional annotation tool 47
Chapter 3 Development of a Burkitt lymphoma classifier 49
3.1 Datasets summary 50
3.2 Choose optimal classification algorithm 51
3.2.1. Algorithms used in previous studies 52
3.2.2. Preparation of the data sets 53
3.2.3. Comparison of different algorithms of multi-classification 53
3.2.4. Comparison of algorithms in binary-classification 57
3.3 Choosing the optimal gene set 60
3.3.1. Find common genes 60
3.3.2. Identify differentially expressed genes 62
3.3.3. Comparison of different gene sets 63
3.4 Cross-platform normalization and training set effect 67
3.4.1. Cross-platform normalization methods 67
3.4.2. Cross-platform normalization effect 69
3.4.3. Training set effects 75
3.5 Conclusion 79
Chapter 4 Validation of the classifier on in-house data from FFPE specimens 81
4.1 FFPE data and preprocessing 83
4.1.1 . Data sets summary 83

4.1.2 . Control probes 84
4.1.3 . Quality check 86
4.1.4 . Preprocessing 88
4.2 Validation of the classifier on the Version_3 and Version_4 datasets 89
4.2.1. Classification Result of Version_3 dataset 89
4.2.2. Classification Result of Version_4 dataset 91
4.2.3. Reproducibility of Version_3 and Version_4 data 93
4.3 Concordance with diagnosis and clinical indication 96
4.3.1. Conventional diagnosis criteria 96
4.3.2. Concordance between GEP classifier and diagnosis 97
4.3.3. Classification of the *MYC*-rearranged DLBCL 100
4.4 R package implementation 101
4.5 Conclusion 104
Chapter 5 The role of *MYC* in non-Burkitt lymphoma 105
5.1 *MYC* translocation expression pattern 105
5.1.1. Datasets summary 106
5.1.2. Differentially expressed genes 107
5.1.3. Predict MYC translocation by gene expression 110
5.2 *MYC* impact on survival 114
5.2.1. Survival impact of GEP predicted *MYC* translocation 114
5.2.2. Survival impact of *MYC* mRNA expression 116
5.2.3. Survival impact of *MYC* mRNA combined with other factors 119
5.3 *MYC* potential mechanism exploration 124
5.3.1. Gene set enrichment analysis 125
5.3.2. Select *MYC*-associated gene lists 128
5.3.3. DAVID functional analysis 130
5.3.4. Potential PRMT5 involvement 132
5.4 Conclusion 133
Chapter 6 Discussion and future work 135
6.1 Discussion 135
6.2 Future work 138

List of References 140

Appendix A The genes in each tested gene lists 155

Appendix B Details of 48 clinically diagnosed DLBCL cases with *MYC* translocation..... 157

Appendix C Detailed GSEA results..... 162

List of Abbreviations

ABC: activated B-cell like lymphoma

aCGH: array comparative genomic hybridization

BCLU: B cell lymphoma with features intermediated between DLBCL and BL

BCR: cell receptor

BL: Burkitt's lymphoma

ChIP-Seq: chromatin immunoprecipitation-Sequencing

CODOX-M/IVAC: cyclophosphamide, oncovin, doxorubicin, and high-dose methotrexate with ifosfamide, vepesid, and high-dose cytarabine

DAVID: Database for Annotation, Visualization, and Integrated Discovery

DLBCL: Diffuse large B-cell lymphoma

DWD: distance weighted discrimination

EBL: endemic Burkitt's lymphoma

ES: enrichment score

FDR: false discovery rate

FF: fresh frozen sample

FFPE: formalin-fixed paraffin-embedded

FISH: fluorescence *in situ* hybridization

FN: false negative

FP: false positive

GC: germinal centre

GCB: germinal centre B-cell like lymphoma

GEO: gene expression omnibus

GEP: Gene expression profiling

GO: gene ontology

GSEA: Gene Set Enrichments Analysis

HIV: human immunodeficiency virus

HMDS: Haematological Malignancy Diagnostic Service

IPI: international prognosis index

KNN: k-nearest neighbours

LDA: linear discriminant analysis

MAQC: microarray quality control

MisgDB: molecular signature database

NHL: non-Hodgkin lymphoma

NIHR: National Institute for Health Research

NMR: nuclear magnetic resonance

PFS: progress free survival

PMBL: primary meditational B cell lymphoma

QN: quantile normalization

R-CHOP: Rituximab Cyclophosphamide, Hydroxydaunomycin, Oncovin, Prednisolone combined treatment

RNA-Seq: RNA sequencing

SAM: significance analysis of microarrays

SBL: sporadic Burkitt's lymphoma

SNP: single nucleotide polymorphism

SVM: support vector machines

TN: true negative

TP: true positive

UCL unclassified DLBCL

VST: variance-stabilizing transformation

WG-DASL: Whole Genome DNA-mediated Annealing, Selection, extension, and Ligation

WHO: World Heath Organization

XPN: cross-platform normalization

List of Tables

Table 1-1: Typical features of BL, DLBCL and BCLU categories 5

Table 2-1: Character variations between two platforms 32

Table 3-1: Datasets summary 51

Table 3-2: Accuracy of 10-fold cross-validation for 10 algorithms in two data sets 59

Table 3-3: Numbers of genes in data sets and used in classifiers 62

Table 3-4: Overall accuracy of tested gene sets in building the classifier 65

Table 3-5: Error rates for classifiers trained on one data set and tested on other public data sets 71

Table 3-6: Accuracy of the classifiers trained by a range of training thresholds 79

Table 4-1: Detail probes comparison between Illumina BeadChips 83

Table 4-2: Number and types of control probes on Version_3 and Version_4 s 84

Table 4-3 : Summary of outliers of each metric 88

Table 4-4: Number of predicted BL and DLBCL in classifiers trained with different training options 92

Table 4-5: BL probability variance of replicates of different normalization methods 94

Table 4-6 : Characters of the diagnosed BL but GEP classified DLBCL cases 98

Table 4-7: diagnosed DLBCL without MYC rearrangement classified as BL 100

Table 5-1: Datasets used in developing *MYC* translocation signatures 106

Table 5-2: Datasets used to validate MYC translocation signatures .. 107

Table 5.3: Number of selected genes under each condition 109

Table 5.4: Classification results of classifiers built with each gene set 110

Table 5-5: Confusion matrix of the *MYC* rearrangement classifier on DASL data 112

Table 5-6: Survival difference between predicted *myc-r* and *myc-n* groups: p-values assessed by Kaplan-Meier model in different treatments and subtypes 115

Table 5-7: Survival difference between various groups: showed in p-values 123

Table 5-8: Additional datasets used in mechanism exploration 124

Table 5-9: GO term annotation of down-regulated genes in *MYC*-rearranged GCB 130

Table 5-10: DAVID annotation for up-regulated genes in *MYC* highly expressed cases 131

Table 5-11: DAVID annotation for *MYC* positive-correlated genes 132

Table: Detailed clinical information of 48 *MYC*-rearranged DLBCL cases 157

List of Figures

Figure 1-1. MYC expression in normal germinal centre cells as well as aberrations leading to different types of lymphoma. 13

Figure 1-2 A simple illustration of microarray technique. 19

Figure 1-3: Work flow of developing and validating GEP Burkitt lymphoma and diffuse large B-cell lymphoma classifier. 25

Figure 1-4: Figure 1-4: Investigations on *MYC*-associated non-Burkitt lymphoma..... 26

Figure 2-1: An example of constructing SVM two-category classification model in a two dimensional space..... 39

Figure 3-1: Prediction results overview of the classifiers built with a list of algorithms and tested on GSE4732_P1 and GSE4475 multi-classification..... 54

Figure 3-2: F-measure for each class of the classifiers built with a list of algorithms and tested on GSE4732_P1, GSE4475 multi-classification. 56

Figure 3-3: F-measure for each class of the classifiers built with a list of algorithms and tested on GSE4732_P1, GSE4475 binary-classification.. 58

Figure 3-4: Venn diagram of genes derived from previous datasets and classifiers. 61

Figure 3-5: Venn diagram of five gene sets used to compare the performance of classifiers. 63

Figure 3-6: F-measure for each class of the classifiers built with a list of gene sets and tested on GSE4732_P1, GSE4475 binary-classification. 65

Figure 3-7: Classification probability correlation of the classifiers built with five gene sets and tested on GSE4732_P1 and GSE4475. 66

Figure 3-8: Violin plot of the gene expression for an example case from each platform when normalized by different methods. 69

Figure 3-9: Prediction results overview of the classifiers when that datasets are normalized by a list of cross-platform normalization methods..... 74

Figure 3-10: Performance of the classifier trained with different BL definition test on GSE4732_P1 with the heatmap of Z-score normalized 28 classifier genes expression value. 76

Figure 3-11: Classification results of GSE4732_P2, GSE10172, GSE17189 and GSE26673 when the classifier was trained with different BL definition plus the heatmap of Z-score normalized 28 classifier genes expression value. 78

Figure 4-1: Illumina Whole-Genome DASL HT Assay combines original DASL Assay and Direct Hybridization Assay. 82

Figure 4-2: Examples of controls probes from Illumina WG-DASL chip..... 85

Figure 4-3: Quality check on each metric of 30 samples from Version_3 chip..... 87

Figure 4-4: Correlation of Version_3 data classification BL probability by different normalization methods and training set options. 90

Figure 4-5: Correlation of Version_4 data classification BL probability by different normalization methods and training set options. 91

Figure 4-6: Classification results of Version_4 data when the classifiers are trained with different BL definition plus the heatmap of Z-score normalized 28 classifier genes expression value. 93

Figure 4-7: Classification reproducibility of the replicates from Version_3 and Version_4 data. 95

Figure 4-8: GEP classification comparison with current clinical diagnosis. 97

Figure 4-9: Number of classic *MYC* rearrangement in different BL probability intervals. 99

Figure 4-10: An example of the classifier implemented in an R package BDC. 103

Figure 5-1: Differentially expressed probed between *MYC*-rearranged and *MYC*-negative non-BLs. 108

Figure 5-2: Venn diagram of differentially expressed probes between *MYC* translocated and *MYC* normal samples in each situation. 109

Figure 5-3: Correlation of clinical FFPE samples *MYC* prediction confidence classified by different lists of genes..... 111

Figure 5-4: Comparison of original diagnosis and *MYC* FISH detection versus GEP BL classification and *MYC* status prediction..... 112

Figure 5-5: *MYC* mRNA expression in FISH detected and predicted *MYC* rearrangement groups of DASL data. 113

Figure 5-6: Kaplan-Meier survival curve in GES 10186 and GSE31312 RCHOP treated cases..... 116

Figure 5-7: Concordance between *MYC* expression and follow up time in each subtype. 117

Figure 5-8: Kaplan-Meier survival curves of four *MYC* expression categories in each dataset. 118

Figure 5-9: Kaplan-Meier survival curves of four *BCL2* expression categories in each dataset. 119

Figure 5-10: Kaplan-Meier survival curves of *MYC* and *BCL2 co*-expression categories in each dataset. 120

Figure 5-11: *MYC* and *BCL2* mRNA expression in different DLBCL subtypes. 122

Figure 5-12: Enriched gene sets in *MYC*-rearranged groups. 126

Figure 5-13: Enriched gene sets in top 20 percent *MYC* high expressed phenotype. 128

Figure 5-14: Mean and standard deviation of the *MYC* expression correlation gene rank lists for all subtypes. 129

Chapter 1
Introduction

The discrimination between Burkitt lymphoma (BL) and diffuse large B-cell lymphoma (DLBCL) has long been a diagnostic difficulty. A group of lymphomas characterized with intermediate features that are hard to assign to one or the other category. In fact these cases have been put into a borderline category “B-cell lymphoma, unclassifiable, with features intermediate between DLBCL and BL” (BCLU) by the World Health Organization Classification (WHO) of Tumours of Hematopoietic and Lymphoid Tissues published in 2008 [1]. While the category probably should not be seen as a new type of lymphoma, rather an approach taken to keep BL and DLBCL as well-defined as possible and to wait for further evidence that allows better classifications in the future. In addition to this highly heterogeneous category, a subgroup of cases that express abnormal MYC activities (genetic and/or expression level), especially when combined with *BCL2* and/or *BCL6* translocations have raised a great attention in recent studies owing to a particularly aggressive clinical course. Even through enormous studies have been reported regarding to the association between MYC and the inferior survival in non-Burkitt lymphomas, the most useful prognostic factor among various MYC activities and the relative significance in the context of other prognostic factors remain controversial, the involved MYC biological function and more proper treatment still require further investigation.

With recent technological advancements, it is now possible to examine genetic abnormalities, mutations and expression on a whole genome scale. This has vastly increased our knowledge of cancer biology, as well as improved the diagnosis and prognosis in the clinics. In the last decade, gene expression profiling (GEP) technology has proven to be effective in the classification of lymphoma subtypes, with several research groups have developed laboratorial level molecular classifiers. Also, whole genome sequencing method has identified genes that are recurrently mutated as well as revealed some oncogenic mechanisms in particular lymphomas. As

promising as it sounds, however it is still a big challenge to make the research findings of clinical use.

Here in this chapter, we first described the lymphomas of interest: Burkitt lymphoma and diffuse large B-cell lymphoma in detail, as well as the role of MYC in those lymphomas, and then we introduced several current popular high-throughput technologies in addition with the criteria of translating research level methods to clinical use. And last we give a general overview of the work carried out in this project.

1.1 Burkitt lymphoma and diffuse large B-cell lymphoma

1.1.1. Burkitt lymphoma

Burkitt's lymphoma is a relatively rare but highly aggressive B-cell non-Hodgkin lymphoma, which contains three clinical variants[2]: endemic Burkitt's lymphoma that is most often observed in African children and associated with Epstein–Barr virus (EBV) infection; sporadic Burkitt's lymphoma that accounts for 1-2% in adults and 30-50% in children of NHL in the United States and Europe [3, 4]; and immunodeficiency-associated Burkitt's lymphoma which refers to cases associated with human immunodeficiency virus (HIV), those occurring in individuals with congenital immunodeficiency, and in allograft recipients. BL is the first human tumour associated with a virus, one of the first tumours that are driven by a chromosomal translocation, and the first lymphoma reported to be associated with HIV infection [5], hence plays an important role in understanding tumour genesis.

All of BL subtypes are similar in morphology, immunophenotype and genetic features [5]. The typical morphology is monotonously uniform and medium-sized neoplastic cells with round nuclei and a "starry sky" pattern resulting from numerous intermixed tangible body macrophages phagocytising apoptotic debris [4, 6]. The immunophenotype involves B-cell-specific and germinal centre associated markers (for example, expression of CD10, CD20 and BCL6), together with the Ki-67 (a protein used as the proliferative index) at nearly 100% (at least >= 95%). The defining molecular feature of Burkitt's lymphoma is the presence of a chromosomal translocation involving

MYC gene and the immunoglobulin heavy chain (IgH) t(8;14)(account for 70~80%) or less commonly *IG* light chain gene t(2;8), t(8;22) (account for 10-~15%) [7], all of which can cause deregulation of *MYC* oncogene.

Burkitt lymphoma was once a narrowly defined NHL but the criteria used to establish a diagnosis of BL has varied somewhat since its original description based on morphologic grounds in its endemic form. Although generally BL is a rather homogeneous group, there are some cases that mimic the morphology or phenotype of DLBCL where diagnosis may be difficult. Currently there is no single character morphology, phenotype or genetic is the gold standard for diagnosis, however it is generally accepted that Burkitt lymphoma usually carries a simple karyotype and encompasses a highly proliferative neoplasm of germinal center phenotype B-cells with deregulation of the *MYC* oncogene in the absence of chromosomal translocations involving oncogenes associated with DLBCL in particular BCL2 and BCL6, although some of the cases have similar characters with DLBCL. Recently the molecular BL (mBL) defined by GEP has received a wide recognition, plus new findings of recurrent mutations including *ID3*, *GNA13, RET,* and *TCF3 (E2A)* [8] are also potentially valuable for BL diagnosis in the future.

Burkitt lymphoma is a highly proliferative malignancy and requires intensive therapy that usually assists with supportive care for toxic effects. Accurate diagnosis is urgent because treatment should be started as soon as possible especially in adults. Now the generally used regimen in UK and USA is CODOX-M/IVAC (cyclophosphamide, oncovin, doxorubicin, and high-dose methotrexate with ifosfamide, vepesid, and high-dose cytarabine) [9, 10]. The outcome for children BL in high-income countries is excellent with an overall cure rate of nearly 90%, while in low-income countries, the outcome are less optimistic which can be caused by incomplete treatment or treatment-related mortality [4]. Outcome in adult patients has been poor, but has improved recently, with a 2-year survival rate about 90% in the young group and about 70% in patients over 65 [3, 5].

1.1.2. Diffuse large B-cell lymphoma

Diffuse large B-cell lymphoma (DLBCL) is the most common lymphoid malignancy in adults accounting for 31% of all non-Hodgkin lymphoma (NHL) in the United States and Europe, with an annual incidence of 7-8 cases per 100,000 people per year [11]. DLBCLs are defined as a group of malignancies composed of large cells with nuclei at least twice the size of a small lymphocyte and usually larger than those of tissue macrophages. Although defined as one group, DLBCL represents a highly heterogeneous type of lymphoma and the fourth edition of WHO classification of Tumours of Hematopoietic and Lymphoid Tissues [1] has further subdivided into morphological variants: centroblastic, immunoblastic, anaplastic, and rare cytologies; molecular subgroups that based on GEP studies : germinal centre B-cell (GCB) like and activated B-cell (ABC) like; immunohistochemistry subgroups: CD5 positive, germinal centre B-cell like and non-germinal centre B-cell like; and other subtypes based on criteria such as the presence of EBV. However, a large number of cases still remain biologically heterogeneous with no clear accepted criteria which subtype should belong to. And these were put into three borderline categories that DLBCL shares intermediate features with other B cell lymphomas that include Burkitt lymphoma.

GEP method has discovered new finding in DLBCL subclassification, the tumours were grouped into two categories that represented different stages of B cell differentiation according to the genes predominately expressed, those associated with germinal B cells that largely expresses genes of normal germinal centre B cells, such as BCL6 and LMO2, and those associated with activated peripheral B cells express genes that are up-regulated in B cells with activated B cell receptor (BCR) signalling, including NF-κB and IRF4. Molecular classifiers based on this cell of origin (COO) discoveries have been used as a gold standard on DLBCL subtypes, and some groups developed an immunohistochemical criteria based on this COO classification for the clinic practical purpose.

Overall DLBCL is aggressive but potentially curable; the standard treatment is R-CHOP – the combination of the Rituximab and CHOP chemotherapy

(drugs used are Cyclophosphamide, Hydroxydaunomycin, Oncovin, and Prednisolone) [12]. The cure rate is variable in different risk groups, ranging from over 80% 5-year progress free survival (PFS) in young patients (< 60) or patients who have a lower international prognosis index (IPI) score, to about 50% in elderly patients or the higher IPI score group. The molecular ABC subtype show generally poor prognosis compared to GCB subtype in patients treated with the previous CHOP regimen as well as the improved R-CHOP treatment [13-15]. This molecular subtype identification has been adopted in a few clinical trials [14, 16], and the difference in the patient outcomes suggests distinct pathways of tumour subtypes that may serve as targets for novel therapeutic strategies.

1.1.3. Cases with features intermediate between BL and DLBCL

There are some aggressive B-cell lymphomas that have the features intermediate between BL and DLBCL, typically between adult sporadic BL and the DLBCL GCB subtype [1, 17, 18]. These cases may have the morphologic features of BL but have greater nuclear and cytoplasmic variability, which is an overlap with the morphologic spectrum of DLBCL [19]. Certain aggressive B-cell lymphomas show Burkitt-like morphology but lack some of the characteristic immunophenotypic findings, or other cases appear to have a classic BL immunophenotype but without *MYC* translocation[20-22]. Besides, the “starry sky” pattern can also be present in a subset of highly proliferative DLBCLs and about 5-10% of DLBCL also carry a *MYC* translocation [10, 23-25], which makes it very difficult to distinguish these two lymphomas by conventional histopathology. The agreement among expert haematopathologists is rather low (~55%)[24].

Table 1-1: Typical features of BL, DLBCL and BCLU categories

Features	**BL**	**BCLU**	**DLBCL**
Cell size	Medium	Medium/Large	Large
Starry sky	Almost always	Very often	Sometimes
Complex karyotype	No	Yes	Yes
Proliferation (Ki67)	>95%	Variable 50~95%	Variable 30~95%

CD10 protein expression	Almost always	Frequent	Variable ~30%
BCL2 protein expression	Negative/Weak	Often strong	Variable
MYC Translocation partner frequency	> 90%, (*IGH*)	35~50%, (*IGL* and *non-IG*)	5~14%,(*IG* and, *non-IG*)
BCL2/BCL6 translocation	No	~45%*BCL2*, ~10%*BCL6*	20~30%*BCL2*,*BCL6*

As stated before the WHO classification has proposed a category named BCLU to summaries the intermediate cases, which usually includes cases have BL morphology but atypical immunophenotypic (eg: BCL2 expression) and genetic (e.g. lack of *MYC* translocation) feature, and cases resemble BL immunophenotypes but show variable morphology (cell and/or nuclear size), and cases present non-*IG MYC* translocation partners with a complex karyotype, or additional translocations of *BCL2* and/or *BCL6* (Table 1-1). The high heterogeneity of this group has confirmed by a continues expression pattern on molecular level, recent analysis of mutation spectra have identified both distinct targets of mutation occurring preferentially in BL including *MYC*, *ID3, TPST2* and *RET,* genes *PIM1, CECR1* and *MYD88* that are predominantly mutated in DLBCL, and genes mutated in both DLBCL and BL with similar frequencies such as *MLL3, TP53* and *LAMA3 [8, 26].* All of which may point to underlying pathogenic mechanisms. However these new finding has not integrated into the diagnostic criteria for routine practice yet.

The diagnose criteria for this intermediate category is proposed mostly due to the feeling that more information need to gather for these biologically heterogeneous cases. However from a clinical point of view it is very important to classify these cases into more appropriate categories, as there is no specific therapy for the intermediate category up till now and we have to choose a more suitable regimen from the existing treatments. In practice BL patients are given intensive chemotherapy, and usually respond rather well if treated in time, however it is not appropriate to give this treatment to

DLBCL patients for the reason it usually very expensive plus toxic (several weeks/months as an in-patient in hospital, expensive drug, can kill the patient or cause morbidity). The majority of DLBCL patients respond fairly well to the standard R-CHOP treatment, which can be given as out-patient appointments, is generally well tolerated, and much less expensive. In clinical reality this biologic diversity has yet not been taken into account sufficiently, current most BCLU cases are given the same treatment as DLBCL, however, several studies found that not all patients are cured by R-CHOP and it is possible that some of the patients currently labelled as DLBCL using standard diagnostic techniques may in fact benefit from the CODOX-M/IVAC treatment [21, 27]. They have indicted that patients with Burkitt-like morphology show a poorer outcome if treated with regimens designed for DLBCL instead of regimens typical for BL. Overall it is still crucial to have a reproducible way to discern the highly heterogeneous intermediate cases as BL or DLBCL correctly and reliably, to reach consensus for optimal management for patients.

Our work on chapter three focuses on developing a robust BL/DLBCL classifier that works effectively across multiple datasets and platforms, and in chapter four we present how the classifier was validated on clinical datasets provided by our collaborators as well as the treatment implication for the intermediate cases.

1.2 *MYC* in BCLU

MYC gene codes for a transcription factor that binds to promoter regions of target genes and modulates their expression by the recruitment of specific activators and repressors [28] [29]. The gene was first discovered in BL patient with the classic chromosomal translocation, however a mutated version of *MYC* is found in many cancers [30]. It is believed that MYC protein regulates approximately 10% to 15% of all human genes [31], and the main functions under its control include, but are not limited to cell proliferation, protein biosynthesis, regulation of metabolism, and the induction of apoptosis [32, 33]. Recent studies suggest that instead of up or down regulating specific groups of genes [34], MYC may target all active

promoters and enhancers in the genome, and acts as a general amplifier of transcription [35, 36].

1.2.1. *MYC* translocation

MYC translocation was initially identified and described as a hallmark in Burkitt lymphoma [37, 38], however it has been recognized in many other non-Hodgkin lymphomas as well , including DLBCL, follicular lymphoma (FL), mantle cell lymphoma (MCL), T-cell lymphoma (T-NHL), plasmablastic lymphoma, and chronic lymphocytic leukaemia (CLL). The true frequency of *MYC* translocation in non-Burkitt B-cell (non-BL) lymphomas is unknown, however it is proved to be higher then previously thought. In contrast to BL, *MYC* translocation in DLBCL is usually found in complex karyotypes and often involves translocation partners other than *IG* genes, also it usually associated with multiple cytogenetic abnormalities, typically concurrent *BCL2* and/or *BCL6* translocations [39-41]. Patients with these combinations of multiple aberrations are reported to have a significantly worse outcome in numerous studies [42-44]. Only approximately 30% of the DLBCL patients who bear *MYC* translocations achieve long-term survival despite modern therapies.

According to the 2008 WHO classification of blood tumours lymphomas with a combination of *MYC* aberration and *BCL2* and/or *BCL6* aberrations are called “double-hit lymphomas” (DHL), while lymphomas with a *MYC* rearrangement but no *BCL2* or *BCL6* rearrangement, irrespective of the presence of other aberrations, are called “single hit lymphomas” (SHL) [1]. As complex as *MYC*-associated non-Burkitt lymphomas can be, a recent study has suggested that there is no obvious difference between double-hit and single-hit aggressive B-cell lymphomas after excluding Burkitt lymphoma [45]. They both have similar frequencies of non-*IG MYC* partner, relatively more complex karyotype compared to BL, and moreover no obvious differentially expressed genes or significant difference in *MYC* expression is observed, suggesting that cases bearing *MYC* translocation in non-BL show some kind of homogeneity.

Therefore it is a reasonable intention to identify an expression pattern specific to *MYC* translocation and see if *MYC*-translocated cases could be

separated from *MYC*-negative cases based on gene expression profiling, and this work is carried out in Section 5.1.

1.2.2. *MYC* prognostic implication

MYC-translocated non-BL accompanied with *BCL2/BCL6* translocations (DHL) is notably reported to have significantly worse outcome [42, 46, 47]. An interesting phenomenon of DHL in DLBCLs is that it is almost exclusively found in the germinal centre B-cell like (GCB) subtype, which generally has a more favourable outcome; it seems the adverse effect of DHL overcomes the better prognosis of GCB subtype. *MYC* translocation is also reported to cause unfavourable outcome, and it is often examined in clinical practice as it is one of the diagnostic marker. However it is not clear that the inferior outcome is due to *MYC* translocation alone or because 58~83% cases also have *BCL2/BLC6* translocations. As *MYC* drawing more attention in the prognosis of non-Burkitt lymphoma, more other form of *MYC* activities are also investigated, with most studies report that *MYC* amplifications are also associated with poor prognosis [43, 47-50], but studies are inconsistent on the association of MYC gains with outcome [8, 51]. Moreover, high MYC mRNA expression [14, 52] and protein expression also have poor prognostic effect even if there is no gene alteration detected [53]. Thus the shorter survival might be caused by various *MYC* activities including genetic, post-transcriptional and post-translational regulation.

It is unclear what the prognostic role of MYC is considering numerous prognostic factors such as age, disease stage, and international prognostic index (IPI) score and other prognostic markers (*BCL2, BCL6*). So far the hypotheses found by different groups are difficult to reconcile and sometimes even contradictory. Some studies report that *MYC* translocation acts as an independent prognostic factor among other well-known factors [44, 54, 55], and factors like *BCL2* translocation do not show significant impact on outcome [44, 55]. While other evidence suggests that *MYC* translocation alone does not cause worse outcome, only when in conjunction with *BCL2, BCL6* or as part of a complex karyotype, does it confer a worse outcome [46, 56]. Or studies showed that *MYC* genetic alterations (translocation, amplification and copy number gains), increased mRNA

levels as well as protein levels are all associated with poor prognosis, but only in the context of concurrent BCL2 protein expression (not *BCL2* translocation) [57]. Recent studies also suggested that B-cell lymphomas with *MYC* and *BCL2* genetic alterations other than translocation behave similarly to DHLs [45], or that it is the co-expression of MYC and BCL2 protein (double expression lymphoma) which contributes to the overall inferior prognosis in both GCB and ABC types [57, 58]. However a problem regarding to the *MYC* and BCL2 protein expression detected by immunohistochemistry (IHC) stains is that it can be very subjective in choosing the cut-off of expression among different labs (e.g. MYC protein expression positive maybe a cut-off on MYC-positive cells with the percentage ranging from 30% to 50%, likewise for BCL2 50%~70%).

In summary, owing to the various forms of *MYC* activities, the rather complex features of most *MYC*-associated non-BLs, other limitations such as technique considerations of detection methods and difficulties in conducting large sample size clinical trials, the prognostic impact of *MYC* still needs further investigation. In this study, we evaluated the survival effect from and mRNA expression angle based on several publicly available large datasets, and the results are illustrated in section 5.2.

1.2.3. *MYC* potential mechanism in B-cell lymphoma

Although various studies have revealed important molecular functions controlled by MYC, including cell growth, protein biosynthesis, introduction of apoptosis, regulation of metabolism and a large number of microRNAs [29, 31, 59], the oncogenic role in aggressive B-cell lymphomas remains largely unknown, especially why it is diagnostic in BL but prognostic in other non-BLs. Part of reason may due to the fact that *MYC* translocation may also present as a secondary change in some lymphomas, complicating a pre-existing abnormality [30, 45].

MYC translocation usually leads to elevated expression on the mRNA and protein levels. An interesting phenomenon is that most of these lymphomas originate in cells that do not normally expression MYC protein. BL and GCB type DLBCL derive from germinal centre (GC) cells, and ABC type DLBCL shows the B cell receptor (BCR) activated features on the plasma cell

differentiation path (see Figure 1-1), in both of which MYC protein is not expressed in normal conditions. Despite the fact that MYC is absent in most GC cells, it is essential for GC formation [37, 45, 60]. The structure of GC contains two parts: a dark zone consisting densely packed proliferating B cells known as centroblasts, where the antigen-driven somatic hypermutation occurs to generate high-affinity antibodies during human immune responses; and a light zone comprised of smaller, non-dividing centrocytes, some of which eventually differentiate into memory B cells or plasma cells [61]. MYC is initially expressed in B cells after interaction with antigens and T cells; however the subsequent up-regulation of BCL6 represses MYC and initiates the formation of the GC dark zone. MYC can be re-expressed in a small subset light zone cells that will re-enter into dark zone after IRF4 up-regulation, and the MYC-absent cells in the light zone will exit the GC as memory cells or early plasmablasts. BLIMP1 induction in these later cells will promote plasma cell differentiation and at the same time repress MYC expression. The expression of MYC and some related proteins in different normal cells are presented in Figure 1-1 with blue colour.

However MYC expression introduced by translocation in the dark zone or maybe by activated BCR signalling in the light zone can contribute to lymphomas, by increasing the proliferation and other metabolic functions. Generally most lymphomas with over expression of MYC show higher proliferation rate, however MYC also introduces cell apoptosis, but it seems that various additional pathogenic mechanisms exist to counteract its pro-apoptotic role and manage to escape cell death [45]. The actual mechanism behind this is still unclear, however recent whole genome sequencing studies have identified several other recurrent somatic mutations, and it seems different types of lymphomas would require various additional aberrations to cooperate with MYC up-regulation. In BL, MYC cooperates with *TCF3* and *ID3* mutations to enhance the proliferation and extend cell survival [8, 50]. The activation of TCF3 and inactivation of ID3 cause a constitutive activation of the PI3K pathway, and promotes the survival of BL cells as well as the proliferation by up-regulating CCND3 [62]. In DLBCLs (or BCLU), it is possible that MYC cooperates with BCL6 and/or BCL2 rearrangements and causes a particularly aggressive type of lymphoma.

Translocation or amplification of *BCL2* can lead to high level expression of BCL2 protein, which has an important function of inhibiting apoptosis. The deregulation of BCL6 also contributes to malignant transformation by increasing the cell growth and lengthening survival time. (Figure 1-1, alterations in red colour). There are many more mechanisms that are pathogenic in these aggressive lymphomas, for example the constitutive activation of NF-κB pathway can also extend cell survival and maybe the reason of poor response to chemotherapy in ABC type.

Moreover cancers arising from the germinal centre usually harbour various mutations/genetic aberrations, and it is extremely difficult to interpret how these alterations work and interact with each other. Nevertheless the question of how MYC regulatory networks interact with other functions remains unanswered. However it is now generally considered that MYC protein itself is “undruggable”, as the pleiotropic role of MYC in developmental biology would prompt the concern of provoking severe, untoward toxicity [29, 63], so potential treatment approaches have been directed at reducing its expression, interfering with MAX dimerization or DNA binding, or acting on the downstream regulatory pathways [64, 65]. The recent discovery that *MYC* transcription depends on the regulatory function of BRD4 has offered new promising therapeutic opportunities[66], and the development of using small molecules targeting the G-quadruplexes formed in *MYC* promoter and MYC/MAX dimerization also provide a promising anti-cancer therapy [67]. In this study we gathered the data and findings have been published previously to explore the potential molecular functions/pathways in non-Burkitt lymphomas, which will be elaborated in Section 5.3.

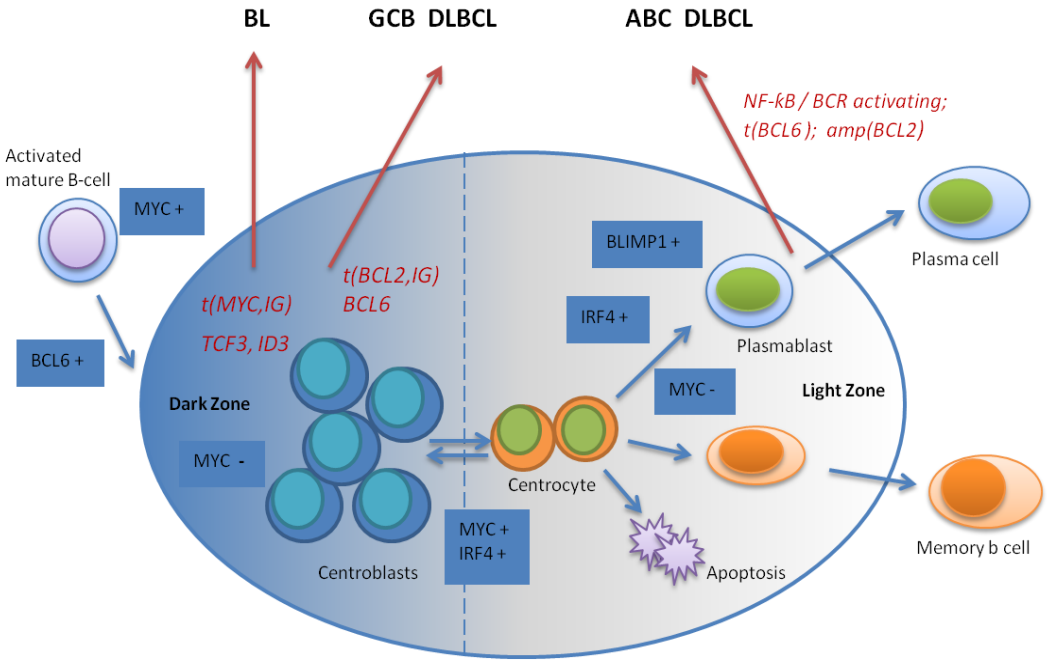


Figure 1-1. MYC expression in normal germinal centre cells as well as aberrations leading to different types of lymphoma.

1.3 High-throughput data in cancer research

Traditional methods for cancer investigation usually focus on one to a few gene alterations or chromosome abnormalities, but with recent technological advances high-throughput measurements have become available in the major 'omics' areas: including array CGH (array comparative genomic hybridization) and SNPs (single nucleotide polymorphisms), array methods to detect genomic variations, DNA microarrays and RNA-Seq (RNA sequencing) that measure gene expression of all transcripts within the whole transcriptome, DNA methylation profiling and ChIP-Seq (chromatin immunoprecipitation-Sequencing) techniques for epignomics analysis, as

well as mass spectroscopy and NMR (nuclear magnetic resonance) spectroscopy technologies used in proteomics and metabolomics studies. Each of the above high-throughput assays offers a distinct perspective of cancer mechanism and potentially provides critical utility in clinic. In next section we give a brief introduction of these techniques.

1.3.1. New methods and techniques

Array CGH is a method that allows the assessment of genetic gains/losses and DNA copy number variations on the whole genome level [68]. It can be used to compare the patient's (cancer) genome against a reference genome and identify differences between the two genomes, hence locating the genetically abnormal regions. It has proven to be specific, sensitive, fast and relatively cheap, and is employed to uncover genetic abnormalities in cancer [69]. A large number of array CGH studies have expanded the knowledge of important copy number aberrations that have crucial clinical use in a variety of tumour types [70-73].

SNP measurements for hundreds of thousands of single nucleotide polymorphism loci spread throughout the genome are promising to explore the common gene variants associated with specific types of cancer [74]. SNPs are quite common and not necessarily causal of disease, also in many cases, SNPs act in unison with other SNPs and with environmental variables, which makes identifying important SNPs difficult. However they have been widely used in genome-wide association studies in discovering new disease loci, SNP associations with drug response and many more aspects [75, 76].

Gene expression levels have influence more directly than gene level mutations, and so significant associations are easier to detect. The revolution in gene microarray has made it possible to simultaneously evaluate the expression level (mRNA) of thousands of genes. And these gene expression profiles are very well suited for classifying patients into cancer subgroups (either better or worse outcomes or with higher or lower values of some phenotypic features) by identifying gene expression levels a subtype is associated with [77, 78]. More recently the advent of next-generation sequencing has made sequence based expression analysis

(RNA-Seq) increasingly popular. Compared to microarrays that only target the genes on the array, RNA-Seq does not require background information and can also look at the expression of transcripts have not been annotated [79, 80]. With fast-evolving experimental protocols, established and computational tools developed, it is likely that this is going to be the favoured method to conduct gene expression pattern analysis in future [81, 82].

Except gene mutations, epigenetic alterations (mainly DNA methylation and histone modification) can also significantly contribute to mis-regulation of gene expression and cause tumour behaviour [83, 84]. Microarray-based or more recent sequencing-based approaches have provided a powerful way to analyse DNA methylation patterns across the genome [85, 86], while ChIP-Seq is a technique for genome-wide profiling of DNA-binding proteins or histone modifications [87]. Both the technologies are indispensable tools for studying epigenetic mechanisms and present a wider picture of epigenetic changes in a cancer genome [88]. The epigenetic findings are entering the clinical field and becoming essential factors in diagnosis, risk assessment, prognosis estimation, therapy management and more aspects [89].

Quantitative proteomics measures are also excellent for the identification of cancer biomarkers that could be used for early detection or classifying people into subgroups, or monitor response to treatment [90]. Mass spectrometry is powerful in identifying the presence or absence of a large number of proteins simultaneously [91, 92]. Advanced techniques in measuring small molecules by mass spectroscopy or NMR [93] have also established the use of metabolomics analysis in cancer research, by detecting metabolic changes in cancer tissues and identifying metabolic biomarkers [94, 95].

Such high-resolution large-scale data types in multiple omics areas definitely improve our understanding of biological mechanisms, oncogenesis and drug effects in different types of cancer, and provide powerful tools for analysis of various purposes. More importantly it is also promising to integrate multiple omics data and conduct system analysis in molecular networks or pathways, ultimately leading to more robust clinical use.

1.3.2. Omics-based tests translation from research to clinic

Although high-throughput technologies have been extensively used to elucidate the biological mechanisms and reveal molecular subtypes or predict prognosis in cancer preclinical studies, only a few of them have been successfully adopted into routine clinical care of patients with cancer [96-98]. Part of the reason is due to the inevitable time delay of translating an initial research test to a well defined and validated test ready for clinical use. However there are many significant challenges lying in conducting an appropriate, robust research design, given the complexity of the generated data, limitations of the techniques or computational approaches, deficiency of plausible biological knowledge, and more importantly criteria and standards in evaluating the reliability in clinic use are strongly needed [99, 100].

The Institute of Medicine of United States has conducted a comprehensive review study by a committee of experts and recommended an evaluation process for determining if omics tests are fit for use in clinical trials [99]. The committee laid out a three-phase process for the development and evaluation of omics-based investigations that aim at clinical use, which are: the discovery phase, the test validation phase, and the evaluation of clinical utility phase.

In the discovery phase omics data of relevant biological/clinical interest are collected then go through quality check and a predictive model development step. In the second stage, built models are validated by a separate dataset, which is usually called a training set and test set validation, where the test dataset is not available an alternative validation called “cross-validation” is performed. However the error rate of the predictive model either by the test set or by cross-validation could be overoptimistic, because there can be similarities in the way the samples were processed. Thus the third phase – an independent clinical dataset is needed to evaluate the clinical utility, which will be for the purpose of the candidate test. It is also important to provide the full descriptions of the independent dataset, as the performance of the model would largely rely on the quality of the dataset. Two “levels of evidence” are reported in the review analysis for the independent dataset

[99]: lower level – clinical data collected at a single institution using carefully controlled protocols, samples from the same patient population, and higher level – data collected at multiple institutions. In which higher level of evidence best assure the test model is less likely over-fitted and robust enough under various situations.

In addition to the suggestions proposed by Institute of Medicine committee, the United States National Cancer Institute working group developed a checklist of criteria for researchers to follow when considering generating an omics-based predictors in clinical trials [100]. A total 30 criteria covering five aspects are listed: specimen issues; assay issues; model development, specification, preliminary performance evaluation; clinical trial design; and finally ethical, legal, and regulatory issues. This paper sets guide lines in assessing the credibility of a predictor as well as pointing out crucial practical issues that must be considered during the test, for instance, specimen quality, amount, collection, processing and storage conditions, technical protocols, reagents and scoring methods used in assay procedures, removal of system effects and normalization in model the development stage.

Therefore our work is carefully carried out under the instructions of above guidelines, considering necessary criteria under each step and full details are presented in following chapters.

1.4 Microarray gene expression profiling

Gene expression profiling (GEP) represents the steady-state level of mRNA under a specific biological condition, which is of great value in understanding cancer biology as well as a mighty tool for cancer classification and prognosis prediction [77, 101-104]. For gene expression represents the functional state of the cancer cell that results from the perturbation of cellular processes whose underlying cause may be mutations or other changes in the genome. DNA microarrays have made gene expression measurements available at whole genome scale, and this especially led to the development of gene expression signatures that may inform prognosis or treatment [105-108]. In addition, tumour behaviour is likely to be dictated in a combinatorial

pattern by abnormal expression of hundreds of genes, which can potentially benefit from global gene expression measurements [109].

1.4.1. DNA microarray technology

The initial description of using complementary DNA microarrays to simultaneously assess the expression levels of thousands genes was introduced by Pat Brown's group in 1995 [110]. Since then the microarrays have been extensively studied for various purposes in cancer research and used as the standard expression analysis platform. The general microarray production process involves taking mRNAs (or total RNA) from tissue samples, reverse-transcribing the mRNA into complementary DNA (cDNA), labelling with fluorescence dye (or radioactive element or other methods) on targets, and hybridization onto probes on a microarray slide (Figure 1.2 give a simple overview of the processing steps). The fluorescent signal is then detected by a scanner, and intensity correlates directly with the relative abundance (usually the intensity of the target probe compare to the intensity of background intensity that is random hybridization) of mRNA present in the sample [110, 111].

The microarray probes can be prepared in different ways; the two main approaches are deposition of DNA fragments and *in situ* synthesis. In deposition-based fabrication, probes are pre-synthesized and then attached to the array by a surface engineering technique. And if it is *in situ* synthesis array, probes are synthesized directly on the arrays using photolithography, ink-jet printing or electrochemical products. Modern microarrays were commercially produced not long after the technology emerged, and several manufactures such as Affymetrix and Illumina are leading the market.

As mentioned above, DNA microarray gene expression profiling has been enthusiastically embraced by cancer research communities. There is no doubt that enormous contributions have made to almost all aspects of cancer research. However there were also several challenges addressed and discussed in the past decade.

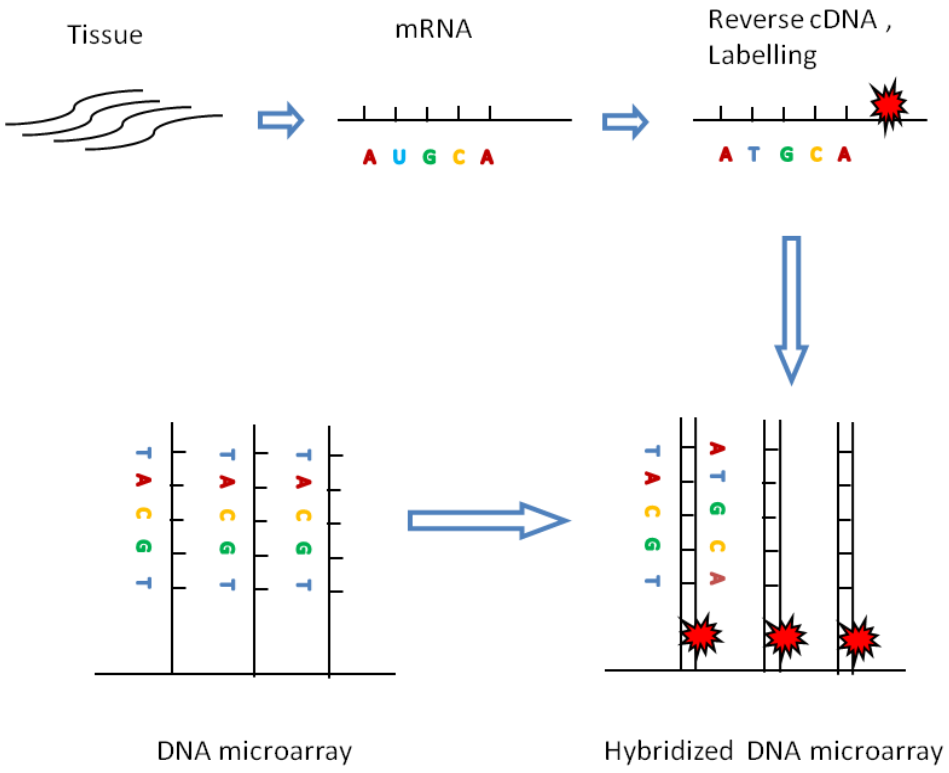


Figure 1-2 A simple illustration of microarray technique.

First is the concern regarding to the hybridization reliability and experiment reproducibility. By the nature of the technique, microarray data tend to be noisy, as the preparation and purification of the targets, along with the hybridization and scanning processes can all lead to differences in the measurements [112]. Thus it is crucial to perform background correction, and a normalization process to remove systematic differences and batch effects. Secondly, as researchers often deposit their data in public databases, this can be then included in other studies as validation or control dataset. However the datasets are usually generated by various manufactures or on different platforms, which have great differences in manufacturing techniques, labelling methods, hybridization protocols, probe length, probe sequence and many other specific features. This raises the worry of whether these data can be combined/compared with similar studies

[113]. Thirdly, apart from the technical obstacles, statistical and computational approaches on DNA microarray are also challenging, because a typical purpose is to find differentially expressed genes among thousands of genes between groups of small sample size, which may cause a number of statistical problems for estimating parameters properly.

These issues are systematically discussed by the microarray quality control (MAQC) project in a series of reports [114, 115]. The conclusions include that it is possible to achieve high intra-platform consistency as well as high level inter-platform concordance as long as the data is processed cautiously and normalized appropriately, and also that model performance depends more largely on the clinical endpoint rather than different approaches in generating the model.

Another problem in validating research findings for clinical application is that, most of the samples investigated in literature are fresh frozen tissue, whereas the generally available samples in routine clinical centres are formalin-fixed paraffin-embedded (FFPE) tissue, from which it is more difficult to extract useful information on DNA microarrays. Recently several research groups demonstrated the improvement of detecting biological signals in FFPE tissues with new platforms and techniques [14, 116]. In summary, molecular classification of cancer subtypes has the potential to become a readily implemented clinical test that may guide future treatment decisions, particularly in identifying those patients most likely to benefit from one of several treatment options just like the situation faced in our study.

1.4.2. General analysis of DNA microarray data

General analysis on DNA microarray GEP is usually divided into two parts: low level analysis and high-level analysis. Low-level analysis focuses on how to get reliable numerical expression data from raw physical data for downstream analysis, and that high level analysis refers to statistical analysis and functional analysis that is related to a specific research question [117, 118]. Below is a brief overview of steps involved in microarray GEP studies, for which the methods applied to each step are explained in more detail in Chapter 2.

Low level analysis mainly includes quality check (QC) and pre-processing. Once the raw intensities have been obtained from the scanning and image processing steps, the quality of each array must be rigorously assessed before undergoing any further analysis. Because even a small number of abnormal arrays can completely compromise the interpretation of microarray data and confound the downstream analysis, thus the outliers should be discarded [119]. On each platform there are certain numbers of control probes used to examine the quality of an array. Preprocessing is a step to eliminate systematic noise and extract meaningful data. It usually consists of three main steps: background correction to subtract background fluorescence signal, normalization to remove systematic bias, and summarization to extract the probe level expression data. [120] There are many preprocessing methods have been suggested as well as various articles been published to compare the performance.

High level analysis usually requires advanced statistic data mining tools to investigate problem of interest such as: identifying the differences between several groups, classifying tumour subtypes, predicting prognosis results or drug response, discovering gene networks and others. Typical analysis involves feature selection, class prediction and mechanism exploration [118, 121]. Feature selection is to find the most differentially expressed genes between groups (normal tissue vs. tumour tissue, various cancer subtypes or samples with different response to treatment, or samples from a series of time points), these genes should be informative to distinguish distinct groups or predict patient prognosis. Class prediction is to apply a mathematical rule on a set of signature genes which is able to predict a new sample into a proper category; there are a wide range of algorithms that have been developed and discussed for this purpose. Mechanism exploration is a further step of GEP analysis involves interpreting the function of interesting signature genes (for example: gene ontology analysis), finding potential pathways or regulation networks or therapeutic targets. This often combines other information like gene ontology terms, pathway databases and useful bioinformatics tools [118, 122].

1.4.3. Reality in DNA microarray GEP analysis

Except understanding the general procedure and crucial aspects that need careful attention in DNA microarray GEP analysis, it is also important to realise the reality: what can and can not a DNA microarray do. Several limitations of this technology exist like: (1) microarray uses relative signal ratio to decide the expression level, and that probes differ in their hybridization properties , thus it's unknown what the true relationship between the hybridization signal measured and the actual RNA concentration is. (2) The use of fluorescently labelled nucleotides to detect the intensity is constrained by the chemistry of the dye especially its sensitivity to oxidation and light [123]. (3) Moreover, microarray requires pre-knowledge of probes and is limited to probes that already have annotation.

Recent advances in next-generation sequencing technologies have introduced a new way, RNA-Seq, to measure RNA expression levels. This quantifies gene expression by sequencing short strands of cDNA, aligning sequences obtained back to the genome or transcriptome, and counting the aligned reads for each gene [124]. RNA-Seq has the ability to identify transcripts that have not been discovered previously and can quantify both very low transcripts (difficult to measure due to background hybridization in microarray) and very high transcripts (may lose because limited amount of probes on the microarray) [125]. Several studies have demonstrated higher sensitivity and reproducibility of RNA-Seq in detecting expression levels [79, 80, 125]. However RNA-Seq is a relatively new technology; it remains necessary to establish appropriate experimental protocols, and problems such as read mapping uncertainty and coverage variation still need to be overcome [82].

In addition, GEP provides only a snapshot of the state of a tumour at the time it is investigated. And the molecular characteristics of a tumour depend on various environmental factors [123]. Therefore, for a better understanding of the cancer biology, combination with other technologies is crucial to improve our ability to achieve the investigation purpose and/or assess the performance of the analysis. However, no doubt that GEP is proven to be a power tool in classifying cancer subtypes.

1.5 Research overview

In section 1.1 to 1.4 we explained the problem we are facing in this project which is classifying BL from diffuse DLBCL as well as the role of *MYC* in those lymphomas, and introduced the background of high-throughout cancer research, especially provided details of DNA microarray expression data that we mainly dealt with. Next in this section we will give a brief overview of the work carried out including: the overall study design and analysis procedures, the investigation environment and developing tools, as well as descriptions regarding data collection and collaborations.

1.5.1. Study design

The first objective of the project is to develop a reliable BL and DLBCL classifier that is ready to use in clinical practice, assisting doctors in the diagnosis and treatment decisions for those intricate cases. The main challenges underlying this study include: (1) lack of optimal classifier, there are competitive classifiers in the literature based on different methodology and gene sets with no clear best choice. Various classifiers have been proposed including methods using flow cytometry cell markers to differentiate BL and CD10+ DLBCL [126], image analysis method that classifying images of lymph sections [127], classifiers that developed based on digital gene expression analysis [48, 49], particularly two GEP classifiers have been widely recognized [128, 129] (2) cross-platform consistency issue, classifiers usually are developed and validated on single dataset that generated from various platforms, and it is not clear whether the classifiers developed or trained on one expression measurement platform can successfully transfer to another, and (3) specimen issue, classifiers developed using fresh frozen samples may not work effectively with the commonly used and more convenient formalin fixed paraffin-embedded (FFPE) samples used in routine diagnosis (usually more noisy).

To conquer those difficulties, we first explore the optimal classification algorithm and gene list applied to a classifier based on the work conducted in two previous GEP studies, and then we test the classifier's reliability and performance cross various platforms with several public datasets by comparing different normalization methods and training set options, then

subsequently we validate the classifier on FFPE samples generated by our collaborators, and evaluate its capability and clinical importance (Figure1-3).

Another objective of this study is to investigate the role of *MYC* in non-Burkitt lymphomas particularly in DLBCL and BCLU subtype. *MYC*-associated non-Burkitt lymphoma has been intensively investigated in recent years, with most studies [27, 42-44, 47, 52, 53, 57, 75, 130-132] reporting that *MYC* abnormalities contribute to adverse clinical course, especially when concurring with *BCL2* and/or *BCL6* translocation. It is no doubt of clinical importance to effectively identify the *MYC*-associated cases and develop trials specifically for this group. However, the design of this type of study is not actually as straightforward as it sounds. Firstly "*MYC*-associated" is a blurry definition that may refer to aberrations at the genetic (e.g. translocation, amplification) level, changes in the level of *MYC* gene expression as mRNA or protein, or acquisition of a '*MYC* associated' global gene expression pattern, and it is unclear which of these is the most relevant [41]. Secondly the *MYC* gene codes for a transcription factor that is believed to regulate expression of 10~15% [133] of all human genes, and plays a significant role in various physiological functions. Moreover recent studies suggest that *MYC* targets different genes in different cancer types [34], which inevitably add complexity to this subject. Thirdly, although the frequency of *MYC*-associated cases might be higher than initially estimated, it still accounts for a rather small percentage of non-Burkitt lymphoma and only till recently the assessment of *MYC* translocation is advised as routine in the clinic [41]. A *MYC*-associated dataset published by a single research group usually contains fewer than 50 aberrant samples, which limits the statistical confidence of conclusions drawn from this type of analysis. Besides, most studies to date focussed on only one single aspect of *MYC* activity, and datasets that contain both genetic (translocation) information and gene expression measurements are extremely scarce.

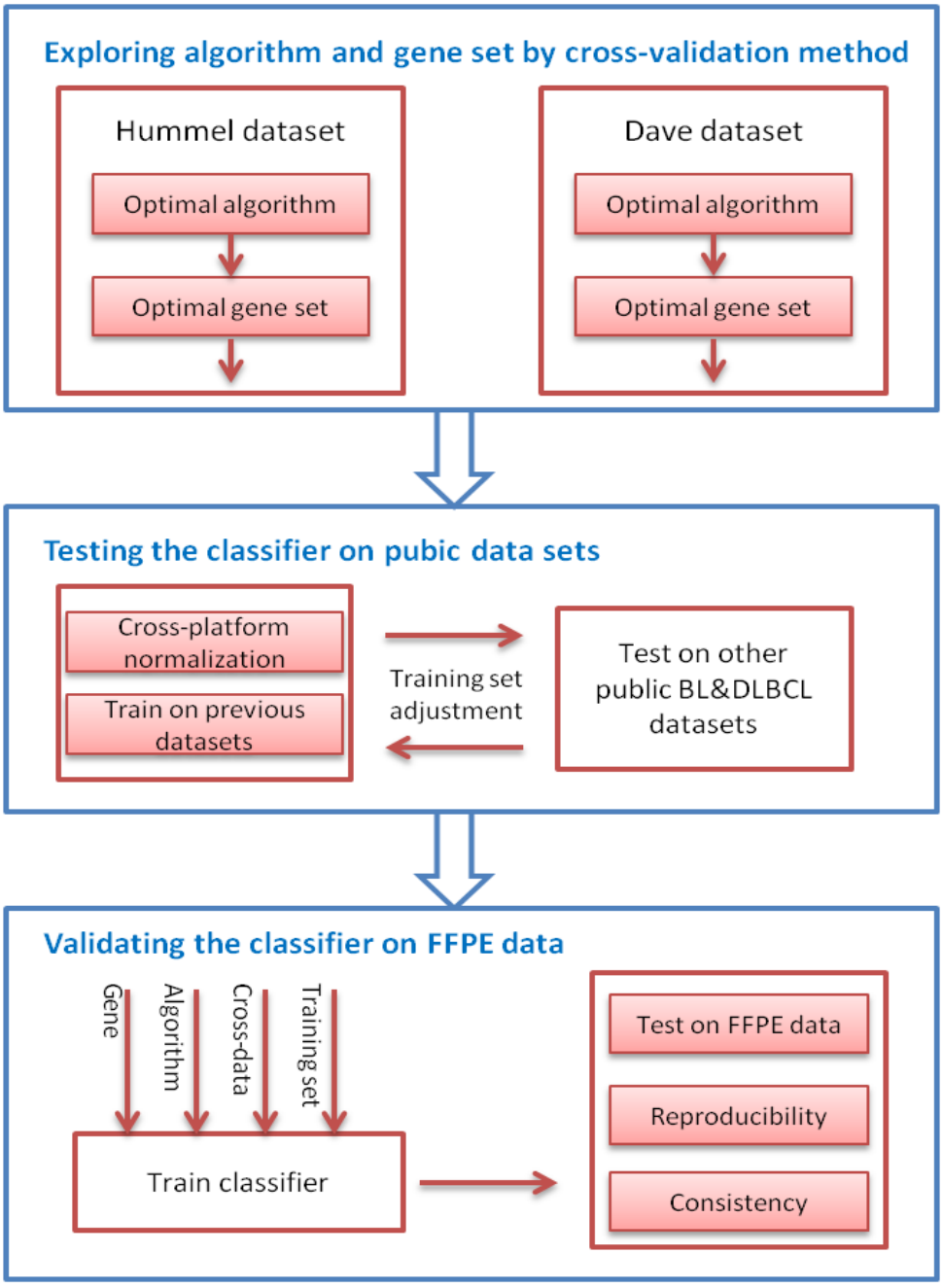


Figure 1-3: Work flow of developing and validating GEP Burkitt lymphoma and diffuse large B-cell lymphoma classifier.

Hence in our study, we collect as many public *MYC*-associated datasets as possible, as well as combining the in-house data generated by our collaborators to investigate the keen problems related with *MYC*-associated non-Burkitt lymphoma from an mRNA expression level. First we try to identify an *MYC* translocation expression pattern, and separated them from the *MYC*-negative non-Burkitt lymphomas by generating a GEP classifier. Second, we assess the survival impact of *MYC* mRNA expression as a single factor also in the context of other factors. In addition, we located gene lists that are potentially MYC-correlated by different means, and then explored the biological functions/pathways might be involved. The results are summarized in chapter 5.

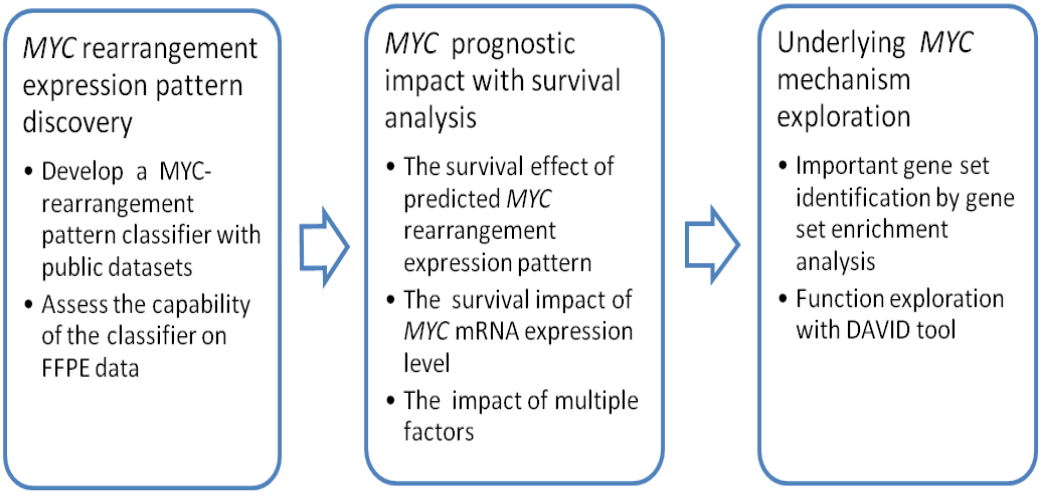


Figure 1-4: Figure 1-4: Investigations on *MYC*-associated non-Burkitt lymphoma.

1.5.2. Developing environment and tools

Most analysis of this study is conducted using R [134]and Bioconductor [135]. R is a language and environment for data manipulation, calculation and graphical display. It can be regarded as a differential implementation of S language developed at Bell Laboratories by John Chambers and

colleagues, and is available as free software under the terms of Free Software Foundation and GNU General Public License, that compiles and runs on a wide range of operating systems. R provides a variety of statistical (linear and nonlinear modelling, classical statistical tests, time-series analysis, classification, clustering) and graphical techniques, moreover it is highly extensible via the package distribution mechanism. There are many packages supplied by R distribution and other useful repositories. The analysis in our study is also implemented into an R package and is shared for public use.

Bioconductor is one of the most useful repositories that provides tools for the analysis and comprehension of high-throughput genomic data, of which most components are distributed as R packages. The Bioconductor project started in 2001 and was an initiative for the collaborative creation of extensible software for computational biology and bioinformatics [135]. It is an open source and open development software with board goals including: provide widespread access to powerful statistical and graphical methods for genomic data analysis; reduce the barriers of interdisciplinary scientific research, and enable the rapid development and deployment of robust, extensible software.

Many of the methods applied in our analysis are implemented in R packages and provided by Bioconductor, there are also some other analysis tool such as Weka machine learning tool, which is open source software in Java that contains a whole range of machine learning algorithms and is well-suited for comparing different algorithms, and Gene Set Enrichment Analysis software developed by the Broad Institute for microarray data analysis, in addition DAVID bioinformatics resources for functional analysis, both of which will be discussed in chapter 2.

1.5.3. Data collection and collaboration

The public available datasets used in this project are downloaded from GEO (gene expression omnibus) database. GEO [136, 137] is a public functional genomics data repository that archives and freely distributes microarray, next-generation sequencing, and other forms of high-throughput functional genomics data submitted by the research community. It provides tools to

help researcher query, locate and download studies and gene expression profiles of interest. It also offers simple submission procedures and formats that support complete and well-annotated data deposits from the research community. GEO stores data in a robust and efficient structure that allows user to locate the platform, research series, and sample record of the expression profiles very easily.

The clinical test datasets are provided by collaborators including Haematological Malignancy Diagnostic Service (HMDS) group in St. James teaching hospital of Leeds, National Institute for Health Research (NIHR) biomedical research centre, and University College London cancer centre. The samples are collected locally from clinical practice as well as from previous/on-going clinical trials. The study has the consent of the participant and is approved by local institutional review board.

Chapter 2
Methods

In last chapter we gave a short introduction to the analysis related with DNA microarray gene expression profiles, and in this chapter we will discuss this topic in more detail, especially concerning the methods applied in the studies reported in this thesis. Section 2.1 explained why low level analysis like quality check and processing are necessary, what can be done to extract meaningful biological information from the experiments, and also how to effectively combine experiments from different studies and/or different platforms by performing cross-platform normalization. Sections 2.2 to section 2.5 give a general overview of several related high level statistical analysis topics (feature selection, classification, survival analysis and functional analysis respectively), as well as the mathematical techniques for a few widely used methods.

2.1 Low level analysis

2.1.1. Quality check

Quality check of the data obtained in a given experiment is absolutely essential, because no valid result can achieved from a compromised array; on the contrary such array usually makes the analysis of other arrays more difficult. In most cases, problems with the quality of an array come from the sample itself. The RNA quality may affect the dye incorporation for spotted arrays and the insufficient RNA concentration may introduce a great deal of spot-to-spot variation. In addition, the mRNA is susceptible to rapid degradation if the sample is not processed properly immediately after collection, and degraded RNA will produce high background noise and low signal intensities on a microarray. Laboratory methods exist and should be used to assess the quality of an mRNA sample, also other laboratory methods are avaible for many of the processing steps involved in the microarray process.

Apart from wet lab quality checks, there are some additional quality control steps can be taken at the low-level data analysis stage. The most straightforward way is to check the detected probes and their intensity distribution. Samples with degraded RNA may lead to overall low intensities. Moreover, each array contains various control probes used to detected different quality problems. These usually include: 1) housekeeping genes that should be expressed in a high level in all samples and could used to detect RNA degraded samples; 2) control probes for hybridization, these are low-stringency, high-stringency, and Cy3 hybridization probes which expect the probes with high-stringency having higher expression then that with low-stringency; 3) negative control probes, which are hundreds of probes of random sequences without targets, these probes should have generally low expression and reflect the background signal from non-specific binding or cross-hybridization. Other types of control probes may also exist depending on the particular array platform. The quality check of arrays is seldom done automatically, because this may be closely related with specific experiments. However there are tools that help to assess control probes and detect outliers.

2.1.2. Preprocessing

Preprocessing is a step that extracts or enhances meaningful expression data, which contains three main parts: background correction, normalization and summarization. The idea behind background correction is that the fluorescence of a spot is the effect of a summation between the fluorescence of the background and the fluorescence due to the labelled targets. We need to subtract the value corresponding to the background: this can be done by estimating the background intensities or by assessing background control probes.

Normalization is a step to remove systematic variation between different arrays, because arrays may have different overall intensities owing to many causes. The goal is to make sure that different arrays can be compared directly and the differences detected between arrays are not introduced by artificial steps. There are several methods to conduct normalization, such as Lowess normalization method [138], piece-wise linear normalization and the

robust multi-array analysis (RMA) method [139] that is popular on Affymetrix data. For Illumina arrays, a variance-stabilizing transformation (VST) [140] is usually applied to take advantage of the larger number of technical replicates. Summarization is the step to generate a single expression for each gene from the probe level data, because a gene can be represented by a set of probes. There are several methods in common use, ranging from simple median or mean value of the probes, or more robust methods considering the hybridization strength variety among probes.

A list of studies has shown that proper preprocessing steps can effectively minimize the artificial interferences, and help to obtain reliable biological conclusions [141, 142]. Usually preprocessing steps are implemented in some R packages and performed together, such as affy [143] package that deals with Affymetrix data and lumi [144] package used for Illumina experiments.

2.1.3. Cross-platform normalization

As researchers often put their data in public database, it would be a great advantage if the data can then be re-used by other researchers investigating similar subjects. And it is also possible to build universal gene expression databases that would compile many different data sets from a variety of experimental conditions. However since modern microarrays are commercially produced, a considerable amount of differences among platforms and manufactures exist: including manufacturing techniques, labelling and detecting methods, hybridization protocols, probe length, probe sequence and numbers used to present certain gene and so on. A simple example of two platforms by Affymetrix and Illumina is showed in Table 2-1.

Heterogeneity of measurement platforms leads to challenges for the re-use of these large data sets, creating limitations for researchers wishing to combine them. More recent studies generally show better cross-platform reproducibility than earlier ones. It is therefore worth asking how data from different platforms might be combined in an analysis. Several cross-platform normalization methods have been developed for the combination of data sets collected using different microarray platforms. These methods are Z-scores, rank scores, quantile normalization (QN) or QN-based methods, and

more sophisticated methods such as the cross-platform normalization (XPN) method [145], and distance weighted discrimination (DWD) [146] by working out the inter-relation among a huge number of genes. Simple methods have the advantage of being easy and fast, while more complicated methods can be effective in complex experiment situations.[147] In our study we compared the performance of Z-score, rank score, XPN and DWD four methods, the results are presented in section 3.4.

Table 2-1: Character variations between two platforms

	Affymetrix	**Illumina**
Platform	HG-U133 plus 2.0	Human HT-12 v4 BeadChip
Labelling	Biotin	Biotin
Probe detection	phycoerythrin-streptavidin-antibody fluorescence	strepatavidin-Cy3 fluorescence
Probe fabrication	In situ photolithography	Pre-synthesized, immobilized on beads, deposited in wells
Probe type	DNA oligonucleotide	DNA oligonucleotide with 29 base address sequence as linker
Probe length	25	50
Probe number	54675	29285

2.2 Feature selection

The low-level analysis output would give the expression profiles of tens of thousands genes (or probes) for a single array, and in common DNA microarray studies there are tens to hundreds of arrays analysed together. To achieve meaningful information from such high-dimensional data requires powerful statistical tools, and that is referred to as high-level analysis. Feature selection is usually the first step of the process. It is essential for biomarker discovery, clustering and classification problems (e.g. cancer subtypes, prognosis categories). Clustering is a type of unsupervised data analysis, in a sense that no pre-assigned class is known but to explore the

expression pattern to see if samples fall into obvious groups (clusters). Two widely used methods are hierarchical clustering [148] and *k*-means clustering [118, 149]. On the other hand classification solves the problem related with building a model that can discriminate pre-existing classes (predict a new sample to a known class). As the classes are known in advance, it is a type of supervised machine learning technique. In this thesis we only discuss the classification problem as it is the major object of our study.

2.2.1. Introduction

Feature selection is a process of deciding a subset of relevant features (in this context probes or genes) that is optimal to discriminate different groups. The main technique is to remove irrelevant (not informative in any context) and redundant (add no more information than current subset) features. The challenge in performing feature selection with DNA microarray data is that the number of features are substantially larger than the number of samples, and it is difficult to identify statistically informative and reliable features as well as to decide the final subset that is best for classification.

Currently there are three main categories of algorithms to address this problem: wrappers, filters and embedded methods [150, 151]. Wrapper methods wrap a particular algorithm and train a predictive model of feature subsets, each subset is then scored by the number of mistakes made on a hold-out set. Wrapper methods usually provide the best performing feature set for that particular algorithm [152]. However, they are very computationally intensive as a new model is trained and tested for each subset, besides the feature subset chosen by wrapper methods can be tuned to the particular algorithm. Filter methods do not involve any classification algorithm, and often produce a ranked list by calculating the correlation, inter/intra class distance or scores of significance tests for each class and feature combinations [113]. Filter methods are more computationally efficient and do not contain the assumptions of a prediction model, so they are more of general use and suitable for exposing the relationships between the features. The weakness of filter methods is that features are considered in isolation, which may lead to a subset that

contains many highly correlated features (redundancy). Embedded methods combine the above, that is to perform feature selection as part of the model construction process, instead of evaluating the performance of each feature subset by counting the mistakes of a predictive model. Other techniques (e.g. shrinking part of coefficients or removing low weight features) can be used to reduce the computational complexity [150].

In our study, we are interested in the differentially expressed genes as well as constructing a classifier to discriminate cancer subtypes, so we conducted feature selection and subtype classification separately. Although different methods may give different results, it is always sensible to find the genes consistently selected by various methods. In the next section we introduced two popular techniques (SAM and moderated t-test respectively) in selecting differentially expressed genes of DNA microarray data, while readers who are interested in other features selection methods can find more details from reference articles.

2.2.2. SAM

SAM (significance analysis of microarrays) is a statistical technique that developed by a Stanford statistic group [153], which is to find significant genes between different groups (normal vs. cancer, different tumour subtypes...). For each gene j it assigns a statistic score d_j taking into consideration the relative change of each gene expression level (difference between means) with respect to the standard deviation of repeated measurement (samples in a specific group). In the context of two classes comparison, d_j is calculated as:

$$d_j = \frac{\bar{x}_{j1} - \bar{x}_{j2}}{s_j + s_0}$$

Where $\bar{x}_{j1}$ is the mean of gene j for class one, and $\bar{x}_{j2}$ is the mean of gene j for class two, and s_j is the standard deviation according to ordinary t-test:

$$s_j = \sqrt{\frac{(n_1 - 1) \times {s_1}^2 + (n_2 - 1) \times {s_2}^2}{n_1 + n_2 - 2}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$$

In which n_1 is the number of samples in class one, and s_1 is the standard deviation of gene j in class1, so are the denotation for n_2 and s_2. s_0 in the

first equation is an estimated variance calculated based on specific problem, the purpose of which is to prevent d_j from being too large when s_j is close to zero (an offset when the standard deviation is biased under smaller sample size).

The significance of the gene is assessed by permutations of the class labels (if null hypothesis is true, there is no difference between groups then the labels should be random). For each permutation p a corresponding statistic score is calculated and ordered $d_{(p1)} \leq d_{(p2)} \leq ... \leq d_{(pg)}$ (g is the total number of genes), and all permutation orders are used to estimate a expected order $\bar{d}_{(j)}$. At this point order the original statistic score as well $d_{(j)}$, the differentially expressed gene are identified by setting different threshold Δ that $d_{(j)} - \bar{d}_{(j)} > \Delta$ (or $< \Delta$ for significantly down). Importantly for each Δ a FDR (false discovery rate) can be calculated from the percentage of significant genes appearing in the permutations statistic scores (if also found $d_{(pj)} - \bar{d}_{(j)} > \Delta$ or $< \Delta$, means the significance is random). Thus SAM provides a way to select differentially expressed gene as well as to assess how reliable the genes are.

This method has more ability than just two groups comparison, and the details can be found in the users guide and technical document [154], also it is implemented in a samr R package for public use.

2.2.3. Smyth moderated t-statistic

Another commonly used method is a moderated t-statistic approach proposed by Smyth et.al [135], which replaces the usual standard deviation of ordinary t-statistic with a posterior residual standard deviation. The model is set up in the context of general linear models for each gene g and comparing groups. Assume in an experiment of n arrays, the expression vector of gth gene is $\boldsymbol{y}_g^T = (y_{g1}, y_{g2}, \dots, y_{gn})$, the linear model is $\boldsymbol{E}(\boldsymbol{y}_g) = \boldsymbol{X}\boldsymbol{\alpha}_g$ with $var(\boldsymbol{y}_g) = \boldsymbol{W}_g\boldsymbol{\sigma}_g^2$, where $\boldsymbol{X}$ is a design matrix, $\boldsymbol{\alpha}_g$ is a coefficient factor, and $\boldsymbol{W}_g$is a known weight matrix. The differences between groups are defined by $\boldsymbol{\beta}_g = \boldsymbol{C}^T\boldsymbol{\alpha}_g$, where $\boldsymbol{C}^T$ is the contrast matrix to represent compared group, and if the null hypothesis is true, for each comparison j, $\boldsymbol{\beta}_{gj}$ equals to zero (one comparison for two groups).

Fitting the linear model to the expression values for each gene will obtain estimators $\hat{\boldsymbol{\alpha}}_g$ of $\boldsymbol{\alpha}_g$, $\boldsymbol{s}_g^2$ of $\boldsymbol{\sigma}_g^2$, the estimators of contrast $\widehat{\boldsymbol{\beta}}_g$ and its variance $var(\widehat{\boldsymbol{\beta}}_g)$can also be derived by the $\boldsymbol{\beta}_g$ definition. There are two assumption on the underlying distribution: (1) the contrast estimators $\widehat{\boldsymbol{\beta}}_g$ are normally distributed; (2) $var(\widehat{\boldsymbol{\beta}}_g)$ and the residual variance $\boldsymbol{s}_g^2$ follow a scaled chi-square distribution. At this point a ordinary t-statistic can be applied with the $\widehat{\boldsymbol{\beta}}_g$ and $var(\widehat{\boldsymbol{\beta}}_g)$**.**

As the same model is fitted to a large number of genes, a hierarchical Bayes' model is set up to take advantage of such information in assessment of differential expressions, and it is assumed the genes are independent of each other. The key to improve statistic tests is to describe how the unknown coefficients $\boldsymbol{\beta}_{gj}$and unknown variances $\boldsymbol{\sigma}_g^2$ vary across genes. This is done by assuming prior distributions for a set of parameters (normal distribution and chi-square distribution). Then the posterior residual variance $\hat{\boldsymbol{s}}_g^2$ is adjusted under the hierarchical model.

$$\hat{\boldsymbol{s}}_g^2 = \frac{d_0\boldsymbol{s}_0^2 + d_g\boldsymbol{s}_g^2}{d_0 + d_g}$$

Of which $\boldsymbol{s}_0^2$ and d_0 are the prior estimated variance and degree of freedom, $d_g$is the degree of freedom for gene g. This way the adjusted $\hat{\boldsymbol{s}}_g^2$ helps to balance the $\boldsymbol{s}_g^2$ when it is estimated from relatively few samples.

This method is available in R limma package, which outputs a p-value toptable for differentially expressed genes.

2.3 Classification methods

Having selected a set of genes that are useful for distinguishing two or more classes of samples, we can then predict the category for new samples based on their expression profiles. The approach of developing a mathematical rule that can effectively decide which category a new sample belongs to based on the existing samples' expression profiles is called classification (or supervised learning in machine learning terminology). Generally classification involves training and testing in two stages. In the

training stage, mathematical rules (classification algorithms) are trained based on the expression profiles of (or part of) existing samples (also called training data), and in testing stage the performance of the algorithm is assessed by samples not used for training (an independent dataset or by hold-out validation or by cross-validation method).

2.3.1. Introduction

There are a wide range of classification algorithms that have been used in GEP analysis,[113, 155] including (1) classical linear methods: Fisher's linear discriminant analysis (LDA), compound covariate predictor [156], logistic regression, Naïve-Bayes classifier, weighted voting; (2) methods based on the class distance or centroid: nearest centroid, shrunken centroid [157], k-nearest neighbours (KNN); (3)decision tree methods and random forests; (4) artificial neural networks: multilayer perceptrons, radial basis function networks, probabilistic neural networks; and (5) support vector machines (SVM).

Here we give a brief description of above classification algorithms. In typical linear classification algorithms, the class label of a sample (or an instance) is viewed as a function of a linear combination of the features (feature vector): $\boldsymbol{F}(\vec{\boldsymbol{x}}_i, k) = \vec{\boldsymbol{w}}_k \cdot \vec{\boldsymbol{x}}_i$, where there are k classes, and $\vec{\boldsymbol{x}}_i$ is the feature vector of i instance, $\vec{\boldsymbol{w}}_k$ is the weight vector for each feature corresponding to class k, the class of a new sample$\vec{\boldsymbol{x}}$ is decided by the value of $\boldsymbol{F}(\vec{\boldsymbol{x}}, k)$. Different algorithms differ in determining the weight vector. For example LDA method assumes the instances in different classes are normally distributed with equal group covariance, hence the weight vector are yield by maximizing means between classes and minimizing variance within each class. KNN method is a nonparametric classification by estimating the class distributions of its k nearest neighbours: this is under the assumption that the characteristics of members of the same class should be similar, and instances located close should be members of the same class.

Decision trees consist of internal nodes and leaves with each leaf representing a class and each node acting as a simple splitter that divides the instances. At each node a cost function is applied to best separate the data (usually measure the "impurity" of the subsamples implicitly defined by

the split), the procedure is performed recursively until some stopping criteria is met.

Artificial neural networks are machine learning methods usually consist of a number of interconnected processing elements (neurons) arranged in layers. The input to the network is the features of an instance and output is the class labels, and there are a few hidden layers in the middle. The output from the input of each layer is implemented in an activation function, and the size and weight of the network are adjusted by the back-propagation algorithm [158] (where forward and backward passes are performed until a stopping criteria is met).

The Support vector machine has raised a great popularity in classification problems, more importantly it is reported to be robust and reliable and usually gives the best performance comparing to other algorithms [113]. SVM is also proved to be the optimal algorithm in our study, and the more detailed introduction is given in following section.

2.3.2. Support vector machines

Support vector machines are learning systems that construct a hyperplane in a high dimensional feature space, which has the largest distance (so-called margin) to the nearest training data point of any class, since in general the larger the margin is the lower the generalization error would be. If the feature space is two dimensional then we are finding a line to best separate the classes, if it is three dimensional then we are finding a plane and the same goes finding a hyperplane in a high dimension feature space. A simple two-dimension two-class example is illustrated in Figure 2.1. Both the green line and the red line can separate the two classes, but the red line has the largest distance (margin) to the nearest points in each class (black circle and triangles in the figures on the margin, also called the support vectors).

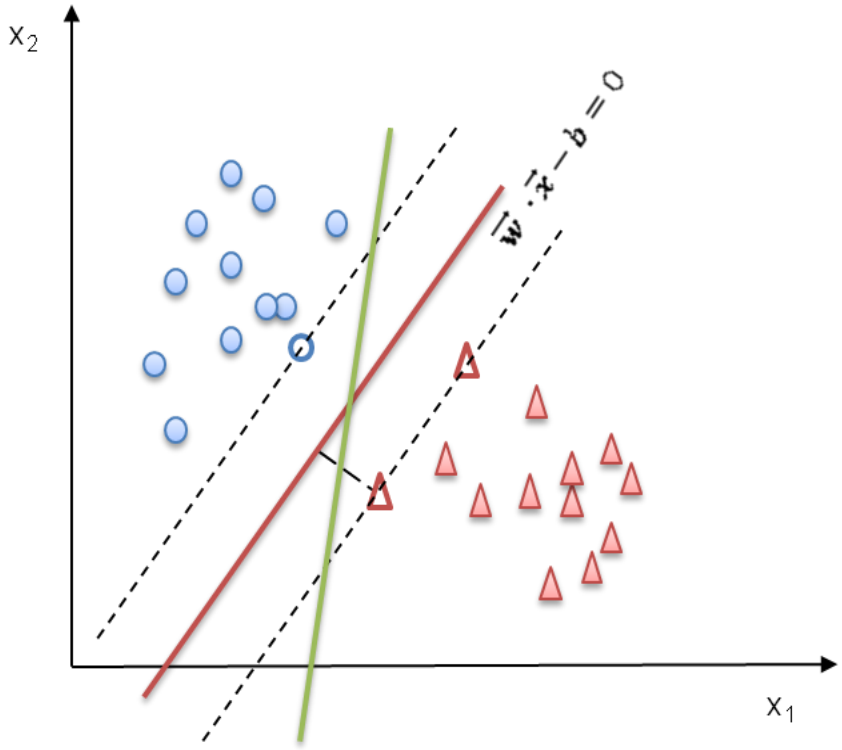


Figure 2-1: An example of constructing SVM two-category classification model in a two dimensional space.

For a set of data points $i = 1,2,\dots,n$, $\vec{x}_i \in R^n$ is the feature vector and $y_i \in \{1,-1\}$ is the class label for each instance. A hyperplane is denoted as $\vec{w} \cdot \vec{x} - b = 0$, where $\vec{x}$ is a set of points in feature space, $\vec{w}$ and b are the weight vector and offset of the hyperplane. The problem of finding the hyperplane is an optimization problem of finding the maximum margin (more commonly finding soft margin when considering mislabelled instances). This is solved by working out the following function:

$$\min_{\vec{w}, b, \xi} \{\frac{1}{2} \|\vec{w}\|^2 + C \sum_i^n \xi_i\}$$

$$\text{subject to } y_i (\vec{w} \cdot \vec{x} - b) \geq 1 - \xi_i, \xi_i \geq 0$$

where C is a constant and ξ_i are a set of non-negative variables called “slack variables”, which measure the degree of misclassifications of the separation.

From an easy understanding, the function can be viewed as a trade off between the first part which maximizing the distance of the margin and the second which minimizing the classification errors.

A great advantage of SVMs is that it applies a dual form representation in solving the optimization problem, which makes it only relevant with the support vectors on the margin, and only the scalar products of the instances needed to be calculated. In addition, the original linear classifier can be easily transformed to a non-linear classifier by replacing the scalar product with so called “kernel trick”. The underlying theory is that the kernel function transforms the feature space to a higher dimensional space, where a hyperplane can be found even if the classes in the original space are not linearly separable. For a kernel function to behave like the scalar product and to ensure the transformed feature space is of geometrical meaning, there are several properties to fulfil (such as the kernel function must be symmetric and follow the Cauchy-Schwarz inequality, also the function needs to follow Mercer’s Theorem, details can be found in reference book [159]). Some common used kernel functions include:

Gaussian radial basis function: $\boldsymbol{K}(\vec{x}_i, \vec{x}_j) = \exp\left(-\gamma \left\|\vec{x}_i - \vec{x}_j\right\|^2\right),\ \gamma > 0$

Polynomial function: $\boldsymbol{K}(\vec{x}_i, \vec{x}_j) = (\gamma \vec{x}_i{}^T \vec{x}_j + r)^d, \gamma > 0$

Sigmoid kernel function: $\boldsymbol{K}(\vec{x}_i, \vec{x}_j) = \tanh(\gamma \vec{x}_i{}^T \vec{x}_j + r)$

The mathematical details are beyond what we discussed in this thesis, and readers who have interests can found references [159-161] useful. SVM algorithms are integrated in a library LibSVM, which can be easily used from other programs.

2.3.3. Evaluation of classifier

Once the features are selected and a mathematical algorithm is taken, a classifier is built for classifying samples into different groups. However inadequate performance caused by insufficient information or overfitting may occur due to a lot of reasons. Whether the classifier serves as a general classification model it can be assessed by a blind validation (an independent data set). However if independent data set is not available, a cross-

validation method (e.g. leave-one-out approach) or hold-out method can be used to estimate the performance of a classifier.

General concepts in evaluating the prediction results are: true positive (TP) and true negative (TN) where outcome classes are the same with pre-labelled classes, while false positive (FP) and false negative (FN) denote the situation where pre-labelled negative but incorrectly classified as positive and vice versa. The detailed number of TP, TN, FP, FN are summarized in confusion matrix. Various performance criteria used including:

$Sensitivity = \frac{TP}{TP+FN}$,

$Specificity = \frac{TN}{TN+FP}$,

$Overall\ classification\ accuracy = \frac{TP+TN}{TP+FN+FP+TN}$,

$Overall\ classification\ error = \frac{FP+FN}{TP+FN+FP+TN}$,

Sometime we also use $F\ measure$ also used to describe $Sensitivity$ and $Specificity$ together.

$$F\ measure = 2 \cdot \frac{Sensitivity\ \cdot Specificity}{Sensitivity + Specificity}$$

2.4 Survival analysis

Another intuitive way to evaluate the performance of a classifier is to compare the treatment response (for instance: remission or relapse) or alive time since diagnosis between predicted classes. If it is known that patients from different cancer subtypes show distinct outcomes and that is also observed in the classified categories, then this could be a persuasive proof that the classifier is able to recognize relevant biological / clinical variants. This type of study is called survival analysis, which examines the probability of event (patient death, relapse) to occur after a fixed period of time. [162] Survival analysis has a widely use in clinical research to evaluate the prognostic impact of a biomarker (which could be a classifier), or to assess the response to a new treatment, or to build a survival model for certain type of patients. It is also introduced into other areas such as engineering (named

reliability analysis), economics (named duration analysis), and sociology (named event history analysis).

2.4.1 Introduction

Survival analysis is a time to an event problem, which attempts to answer questions like: what is the probability that a patient is still alive after a certain time? Is there a survival probability difference between two groups of patients? What are the factors that affect the survival probability and further to what extent the factors increase or decrease the probability? This is usually measured by monitoring the survival time of a large patient population under different conditions. However, there are several difficulties in performing this type of analysis. First it is usual that up some patients may not have the event of interest at the end of the monitoring time (follow-up), and thus their true time to event is unknown. Further, not like in other statistic analysis, the survival time is complex and seldom normally distributed, while the survival data of patients are usually skewed to many early events and relatively few late ones. [163]

Before introducing several analysis methods, we first explain some basic concepts. Survival data are generally described in terms of two probabilities, survival and hazard. Survival probability (also called survival function) $S(t)$ is the probability that an individual survives from the start point of the study (diagnose time) to a specified time t. The hazard is usually denoted by $h(t)$ and is the probability density that an individual who is under observation has the event at time t. [163] Thus the survival is the cumulative probability of non-occurrence until certain time, while hazard reflects the event occurring rate; it is important because this provides insight into the conditional incident rate and a approach to form a survival model.

Another important concept is called censoring, which deals with the difficulty that the actual event time of an individual might not be known. This may caused by: (1) an observed individual hasn't experienced death by the time of the study closed; (2) individual is lost to follow-up during the study period; (3) individual suffers another event which makes further follow-up impossible. This situation is often called right censoring, and is the most common type. There are also left censored and interval censored type of

data, which is less common and not illustrated here. The import assumption is that censoring is not related to the probability of an event occurring (does not affect the occurrence rate). Although none of these subjects suffer the event of interest, the fact that they contribute time without suffering an event is vitally important [162].

2.4.2 Kaplan-Meier survival estimate

Kaplan-Meier method is a simple nonparametric estimator of survival data that doesn't make any survival distribution assumption, and is one of the most widely used techniques [164]. This method is used to estimate the proportion of people who survive after a specific time point. The survival function is calculated based on the number of surviving patients at each point and the cumulative number of events have occurred up to that point. For a sample of size k, let the observed times be $t_0 < t_1 < t_2 < (\dots) < t_k$. The probability of being alive at time t_j is a conditional probability on t_{j-1}, n_j is the number of individuals alive just before t_j, and d_j is the number of event at time t_j, then the survival function is described below:

$$S(0) = 1,\ S(\infty) = 0$$

$$S(t_j) = S(t_{j-1})\left(1 - \frac{d_j}{n_j}\right)$$

This is very easy to understand that the probability of an individual will survive till a given time is the cumulative probability of survival since the beginning and the probability of survival drops only when an event occurs. The calculation and presentation of the Kaplan–Meier estimates are usually displayed in a life table; more often a survival curve (survival plot) is used to present the data. Especially the survival curve can be used to compare survival differences between different groups. Also the significance of the survival difference can be checked by statistic methods, among which log-rank test [165] is most commonly applied [163].

Whilst this technique is easy to use and interpret, it has its limitations. Although differences between groups can be seen and their statistical significance also can be tested, no estimate of the actual effect size is quantified as it makes no assumptions about the underlying distribution and

no attempt to describe the effect numerically. Also more often we are interested in the survival effects of multiple factors (multivariate / covariate), where this approach is limit by single factor. In addition where there are imbalances between groups, the findings will be prone to confounding and bias. Moreover the log-rank test is only applicable to categorical effect factors, which makes it rather limit on continuous factors.

2.4.3 The cox proportional hazard model

There are several parametric models that make assumptions about the distribution (shape) of the hazard function and the impact of variables on survival data, such as exponential model assumes the hazard is constant over time, while Weibull and Gompertz models assume the hazard is always increasing or decreasing as time forwards, and Log-Logistic assume the hazard either rises to a peak point then decrease or always decreases.[163] The estimated shape of the hazard is modelled by responding parameters. However it is rather difficult to choose a model that is suitable for a certain scenario, and the most widely used technique at present is a semi-parametric model developed by Sir David Cox [166] called the Cox proportional hazard model, which makes a statistical assumption on the effect factors (covariates) but no assumption on the underlying shape of the hazard. The model is written as

$$h(t) = h(t_0) \times exp\{b_1 x_1 + b_2 x_2 + \cdots + b_p x_p\}$$

where the hazard function $h(t)$ depends on the baseline hazard function $h(t_0)$ and a set of covariates $(x_1,\ x_2, \dots, x_p)$. The impact of the covariates is measured by coefficients$(b_1, b_2, \dots, b_p)$, and it is assumed to be proportional that is the hazard of the event in any group is a constant multiple of the hazard in any other. The quantities $exp(b_i)$ are called hazard ratios, and give a way to quantify the covariates impact, which a hazard ratio is greater than 1 indicates a covariate is positively associated with hazard function (event occurrence probability).

The validity of the proportional hazard assumption could be test by “log-minus-log” plot or log(−log(survival)) plot, which is the logarithm of the cumulative hazard function in each group against the logarithm of time, and

it should give lines that are parallel [163]. Also the covariates and interactions could be tested by adding interaction terms.

2.5 Mechanism analysis

The ultimate goal of gene expression analysis (or any type of bioinformatics analysis) is to understand and describe the basic biology questions, such as what are the molecular processes and pathways related to a certain cancer (disease), and this is usually explored by explaining the meaning of the genes that are chosen with specific standard (e.g. differentially expressed between two phenotypes). The Gene Ontology (GO) Consortium [167, 168] developed three structured, precisely defined, controlled vocabularies (ontologies) to describe gene and gene products in terms of their associated biological processes, cellular components and molecular functions in a species-independent manner. Each node in the GO ontologies are linked to other kinds of information, including the many gene and protein keyword databases such as SwissPROT [169, 170], GeneBank [171], PIR [172] to keep up with the rapid changing of these biological knowledge.

Therefore a standard strategy for interpreting the biological meaning of a list of genes derived from high throughout experiments usually conducted by first mapping the genes to associated Gene Ontology (GO) terms, and then statistically highlighting the most over-represented (enriched) GO terms in the list. Enrichment is a promising analysis that increases the likelihood for investigators to identify biological processes most pertinent to the biological phenomena under study. There are a number of similar public available tools for this type of analysis, including but not limited to GOMiner [173], GOstat [174], GFINDer [175], GSEA [176] and DAVID [177]. Here in our study, we applied two currently commonly used tool GSEA and DAVID, and a brief introduction is given in the following section.

2.5.1. Gene set enrichment analysis tool

Early methods for gene expression analysis focus on finding a set of genes that are differentially expressed between distinct phenotypes. However there are several limitations such as being difficult to interpret the biological meanings, and where to draw the cut-off for the number of genes that are

significant, also it is believed that some disease related genes may have modest changes. Hence methods based on gene sets are proposed to take more use of the gene expression data. The gene sets are those have similar molecular function, or involved in the same pathway or known to be coexpression from previous experiments. A particular method named Gene Set Enrichments Analysis (GSEA) [176] was developed by Subramanian et al in 2005.

The basis strategy is first rank all genes by the magnitudes of their differential expression into a list L, and then determine whether gene members from a gene set S tend to occur toward the top (or bottom) of the list L (so called enriched). For each gene set S, an enrichment score (ES) is calculated according to a weighted Kolmogorov–Smirnov-like statistic (not explained here), and then the significance level of the ES (nominal p-value) is estimated using an empirical phenotype-based permutation test. Finally the ES is adjusted by multiple hypothesis test to yield a normalized ES (NES), and the false discovery rate (FDR) corresponding to each NES is computed to evaluate the significance of that specific gene set.

GSEA tool takes two main files as input: a file containing expression values of all the genes and another file indicating the phenotype of the cases to be compared. The enrichment score is evaluated using a collection of gene sets repository MSigDB Molecular Signatures Database[178], which is one of the most widely used knowledgebase containing annotated sets of genes involved in biochemical pathways, signalling cascades, expression profiles from research publications, and other biological concepts. Users can choose the gene sets from 7 available collections as well as provide their own gene sets to run the enrichment analysis.

GSEA presents a number of benefits in gene expression data analysis, for it focus on gene sets rather than significant genes, where the gene sets may have a biological lead to potentially involved molecular functions, pathways or regulatory networks and others. This is an effective way of linking prior knowledge and uncovering interesting new findings.

2.5.2. DAVID functional annotation tool

DAVID (Database for Annotation, Visualization, and Integrated Discovery) [177] provides a set of data mining tools that systematically combine functionally descriptive data with graphical displays to promote discovery through functional classification, biochemical pathway maps, and conserved protein domain architectures. The DAVID bioinformatics resources consists of an integrated biological knowledgebase built around the DAVID Gene Concept, which is a single-linkage method enabling a variety of publicly available functional annotation sources to be comprehensively integrated and centralized by the DAVID gene clusters [179]. Three main functional analytical tools [180] are:

Functional annotation chart that provides typical gene term enrichment analysis, to identify the most relevant (overrepresented) biological terms associated with a given gene list. DAVID owns extended annotation coverage over 40 annotation categories, including GO terms, protein–protein interactions, protein functional domains, disease associations, bio-pathways, sequence features, homology, gene functional summaries, gene tissue expression and literature [179]. The annotation categories can be flexibly included or excluded from the analysis on the basis of a user's choices. In addition, users can define their own gene population backgrounds and tailor the enrichment analysis to meet specific analytic situation.

Functional annotation clustering that provides the ability to explore and view functionally related genes together as a unit, so that to concentrate on the larger biological network rather than at the level of an individual gene. Currently the majority of co-functioning genes may have diversified names and that genes cannot be simply classified into functional groups according to their names. Hence a set of novel fuzzy clustering techniques is adopted to cluster the genes that are somewhat heterogeneous, yet highly similar annotation into functional annotation groups. This type of grouping of functional annotation is able to give a more insightful view of the relationships between annotation categories and terms compared with the traditional linear list of enriched terms.

Another analysis tool is functional annotation table, which is a query engine for the DAVID knowledgebase, without statistical calculations. For a given gene list, the tool provides the corresponding annotation for each gene and present them in a table format. This is useful especially when investigators want to closely look at the annotation of certain highly interesting genes.

It is a popular function analysis tool and has been widely used by other researchers, here we use DAVID to investigate the biological mechanisms of *MYC* in lymphomas.

Chapter 3
Development of a Burkitt lymphoma classifier

Many efforts have been made to distinguish Burkitt lymphoma and diffuse large B-cell lymphoma due to its clinical importance, especially on the cases where conventional diagnosis is difficult. And the successful use of gene expression profiling in classifying patients into different cancer subtypes has stimulated researchers to solve this problem from molecular level. Typical approaches of developing a GEP classifier includes first finding a set of most differentially expressed genes between pre-assigned BL and DLBCL groups, which were usually diagnosed by clinicians according to certain criteria (e.g. morphology, phenotype), and then adapting an appropriate classification rule on the gene set so that it can predict which category a new case belongs to.

Two studies performed by Dave et al. [128] and Hummel et al. [129] respectively have done excellent work related to this aspect: each identified a set of signature genes and applied a classification method that can accurately recognize BL from DLBCL. Although both studies have established a molecular definition of Burkitt lymphoma and successfully distinguished it from DLBCL, there are vast differences in the developing stages including gene sets and methodologies applied in the classifier, and it is not clear which is a better option and to what extent the classifiers agree with each other. In addition, both classifiers were cross-validated on their own developing datasets only, thus it is not known how they would work on independent datasets generated by other groups and/or on different platforms.

In order to develop a robust classifier that is able to assign samples into correct lymphoma categories, despite of the platforms where the samples are generated from, we performed a comprehensive comparative analysis in this study. First a variety of classification algorithms and gene sets were thoroughly investigated so as to build a classifier that best recapitulates the classification results from the previous studies. Then the transferability of the new classifier among datasets on different platforms was assessed with a

series of cross-platform normalization methods. Finally we compared the stringency of BL definition adopted in the two published classifiers and carefully adjusted the training set for the new classifier.

3.1 Datasets summary

Burkitt lymphoma is a relatively rare type of non-Hodgkin lymphoma and the gene expression data available is rather scarce. A total of six datasets were used in this chapter, all of which were downloaded from the Gene Expression Omnibus (GEO). Two of them were generated by the groups that developed the previous classifiers, and here were used as the development datasets to examine the algorithm and gene set options for the new classifier. The other four datasets were used to test the classifier's transferability on various platforms. The data sets used in this chapter are summarized in Table 3-1.

GSE4732_P1 is the data set produced by the Dave group [128] on a custom oligonucleotide Affymetrix microarray, with 2524 unique genes that are expressed most differently among various types of non-Hodgkin lymphomas. There are 303 samples in the dataset, which were classified into 54 BL and 249 DLBCL subtypes: 91 ABC, 95 GCB, 30 UCL also 33 PMBL (primary mediastinal B cell lymphoma). The Hummel dataset [129] GSE4475 was performed on the Affymetrix U133A Gene Chip and a total of 221 samples (containing one replicate sample) were classified into 44 mBL (molecular Burkitt lymphoma) and 177 non-mBL plus 48 intermediate cases.

The additional four validation datasets are: a subset of the samples from the Dave dataset performed on another platform containing 33 BLs and 66 DLBCLs, here referred as GSE4732_P2; GSE17189 [181] which is an HIV-related dataset that consists of 14 HIV-related DLBCLs and 3 HIV-related BLs; and 15 samples from GSE26673 [182] that include 13 endemic BLs (EBL) and 2 HIV-related DLBCLs. All the above three datasets were generated from experiments on the Affymetrix HG-U133 plus2.0 Gene Chip. The last dataset, GSE10172 [183], from the Affymetrix HG-U133A Gene Chip contains 36 BL and paediatric DLBCL samples.

Table 3-1: Datasets summary

GEO No.	Group	Sample	Probe	Platform
GSE4732_P1	Dave	54 BL, 249 DLBCL	2745	Custom Affymetrix Lympho-Chip
GSE4732_P2	Dave	33 BL and 66 DLBCL subset of GSE4732_P1	54675	Affymetrix HG-U133 plus2.0
GSE4475	Hummel	44 mBL, 48 intermediate, 129 non-mBL	22283	Affymetrix HG-U133A
GSE10172	Klapper	13 mBL, 9 intermediate, 14 non-mBL	22283	Affymetrix HG-U133A
GSE26673	Piccaluga	13 EBL, 1 sBL, 2 HIV-BL	54675	Affymetrix HG-U133 plus2.0
GSE17189	Deffenbacher	4 HIV-BL, 13 HIV-DLBCL	54675	Affymetrix HG-U133 plus2.0

3.2 Choose optimal classification algorithm

Classification on microarray expression data is a challenging task due to the typical high number of genes and small number of samples. A well performing method should learn a classification model from the training set and further predict new samples into certain class under a low error rate. Popular methods used in gene expression data include Bayes classifiers, classification trees, neural networks, support vector machines and some other discriminate analysis schemes (this is also discussed in chapter 2). Although many methods have been proposed and new methods are continuingly being developed, most of them were only applied in one single dataset, and thus the advantage against other methods is somewhat less convincing. In addition, it is difficult to find an algorithm that works best in all cancer type classification problems, and the choice of gene selection criteria also has effect on performance. In our work we considered a list of well-acknowledged algorithms in the context of classifying BL from DLBCL using

two separate datasets based on the gene signatures developed by previous studies.

3.2.1. Algorithms used in previous studies

Both of the previous studies designed their own method to classify BL and DLBCL. Dave group built a two-stage three pair-wise Bayesian compound covariate predictor, which the samples are classified to a category by different gene sets used in the two stages, and three pair-wise predictions are BL against three DLBCL subtypes: ABC, GCB and PMBL respectively. The genes used in each stage are: (1) *MYC*-target genes identified by RNA interference, (2) 100 genes that have the most significant t-statistic between the pathologists agreed BL and each DLBCL subtype. For each sample, first a linear predictor score was calculated, then Bayes rule was applied to estimate the probability of belonging to one of the two categories, and a sample is classified as BL only when it is predicted as BL in both stages of the predictor and in each of the three pair-wise comparisons.

On the other hand Hummel and colleagues defined a molecular BL class by implementing a core group extension method. They used 8 classic BLs that met all World Health Organization (WHO) criteria as core group to generate a molecular Burkitt lymphoma signature index, and calculated the posterior probability of the other samples belonging to the core group. If the probability of BL was over 0.95, the sample was assigned to molecular BL named mBL, and if it is less than 0.05 the sample was assigned to non-mBL; the samples with probability in between were summarized in an intermediate class.

While these two methods represent useful developments, it is not clear that they are the best possible algorithmic choices. Here we compared 10 widely known algorithms: Naïve Bayes, Bayes Net, LibSVM, SMO (sequential minimal optimization), Neural Network, RF (random forest), FT (function tree), LMT (logistic model tree), REP (reduced-error pruning) Tree and J48 pruned tree, for their performance on predicting above two datasets.

Generally classification algorithms are evaluated by the agreement percentage with the high confidence samples diagnosed by experienced experts. However, in this study the algorithms were evaluated according to

the original assigned class by previous classifiers, with the simple aim of generating methods able to recapitulate previous results.

3.2.2. Preparation of the data sets

The Hummel dataset was processed the with affy package [143] from raw data and expression summarization done with the rma algorithm [139] with quantile normalization. Then the 58 classifier genes consisting of 74 probes picked in the original classifier were extracted and put into Weka machine learning tool. The Dave dataset did not use a standard Affymetrix array and the raw data cannot be processed with the affy package; hence the text format data which contain the expression values that have been preprocessed and normalized by the author already was used in the analysis. Of the 217 genes used in Dave's classifier, there were 9 genes for which we were unable to locate the correspondent probes according to the author's annotation file; further mapping on the HGNC website identified another 6 genes from the dataset, so finally 214 genes corresponding to 234 probes were extracted and put into Weka for algorithm comparison.

3.2.3. Comparison of different algorithms of multi-classification

The classifiers built by various algorithms were examined both on multi-classification (5 classes in the Dave group and 3 classes in the Hummel group) and binary-classification (BL compare to DLBCL). We used the algorithm implemented in Weka version 3.6, with all parameters set to default except giving the number of initial trees for RF as 100. All classifiers were evaluated by 10-fold cross-validation within each dataset, and the samples predicted differently from the original class were regarded as mis-classified cases.

The general overview of the classification results by each method on dataset GSE4732_P1 and GSE4475 is shown in Figure 3-1 A and B respectively. In both datasets most of the algorithms can separate BL from others with minor differences.

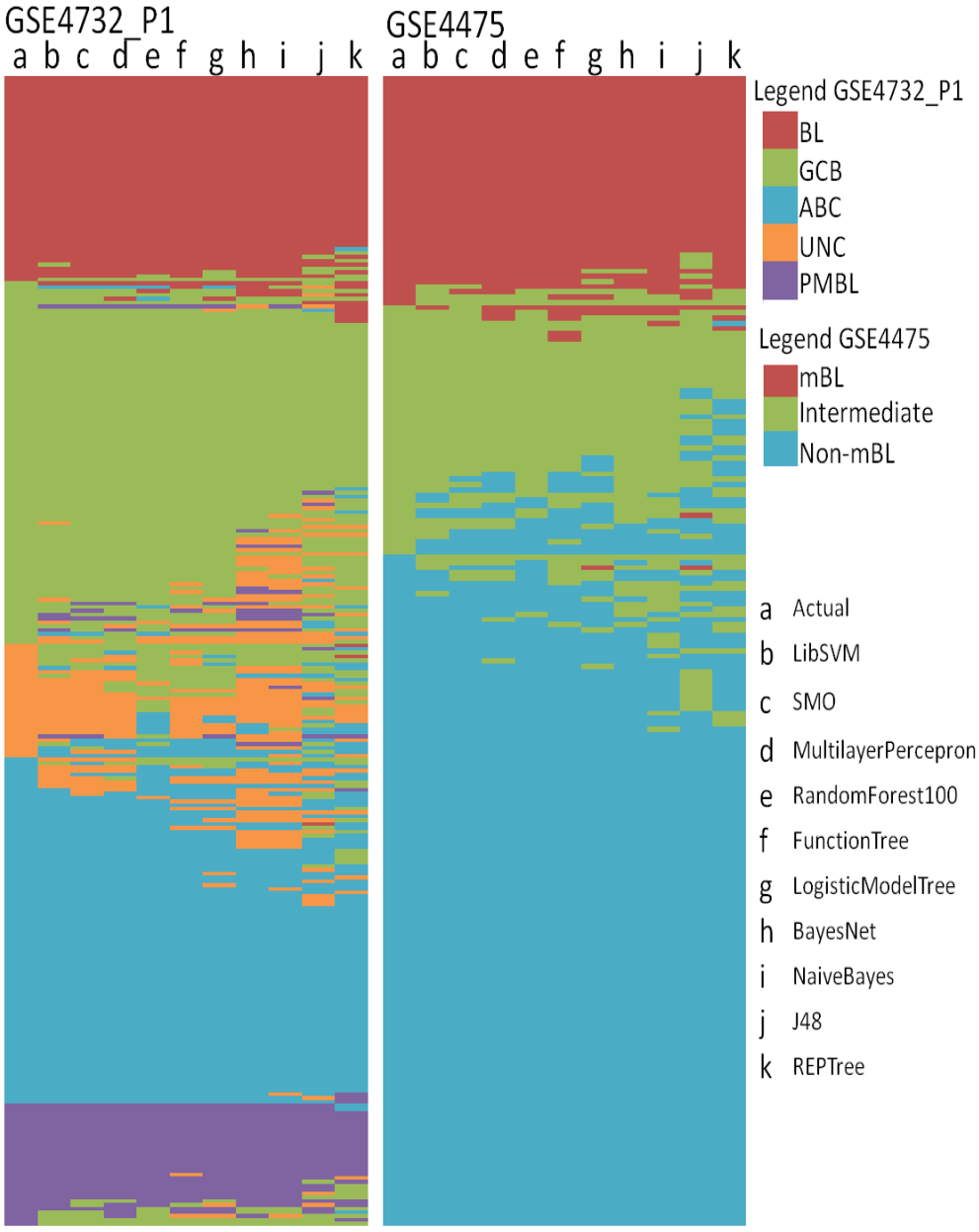


Figure 3-1: Prediction results overview of the classifiers built with a list of algorithms and tested on GSE4732_P1 and GSE4475 multi-classification.

However there are clear differences in the ability of the algorithms on categories besides BL. The last two methods J48 and REP Tree have difficulty distinguishing DLBCL subtypes in the GSE4732_P1 dataset or the intermediate cases in dataset GSE4475. The two Bayes methods tend to assign more cases as unclassified in GSE4732_P1 or intermediate in GSE4475, while the three tree methods, especially RF, seem to classify the unclassified cases into ABC and GCB type in GSE4732_P1 and some intermediate case as non-mBL. LibSVM, SMO and Multilayer-Perceptron gave the classification more close to the actual class (classes defined by previous studies).

The detailed performances of all classifiers are compared in Figure 3-2. In multi-classification, the overall accuracy does not give much information on how the classifier performs on each class; hence the F-measure of each class is used to assess the capability of the algorithms. F-measure considers both the sensitivity and the specificity of a particular class, and it equals to 1 when the classifier recognizes all the cases belong to this class meanwhile all the cases classified into this class are actually originate from the class.

The result of dataset GSE4732_P1 shows that most algorithms can accurately distinguish BL from all DLBCL categories. While the algorithms perform less effectively within DLBCL subtypes, and the F-measures of different algorithms vary widely. However LibSVM and Multilayer-Perceptron methods still work rather well with an F-measure over 0.85 in all classes. When we tested the algorithms on dataset GSE4475, similarly almost all algorithms except J48 and REPtree can classify mBL and non-mBL cases with F-measures close to 0.95, but again the performance drops significantly on the intermediate cases. This could mainly be because the intermediate category is rather heterogeneous, as it contains the cases where the authors could not confidently assign as BL or DLBCL. Nevertheless we can see that LibSVM, SMO showed better results than other algorithms with F-measure over 0.8.

To summarize the results from both datasets, under default parameters the performance of the above algorithms differ significantly by 10-fold cross-validation. J48 and REPtree gave the least satisfactory results, while

LibSVM clearly built the most trust worthy classifier that reproduces the original classes.

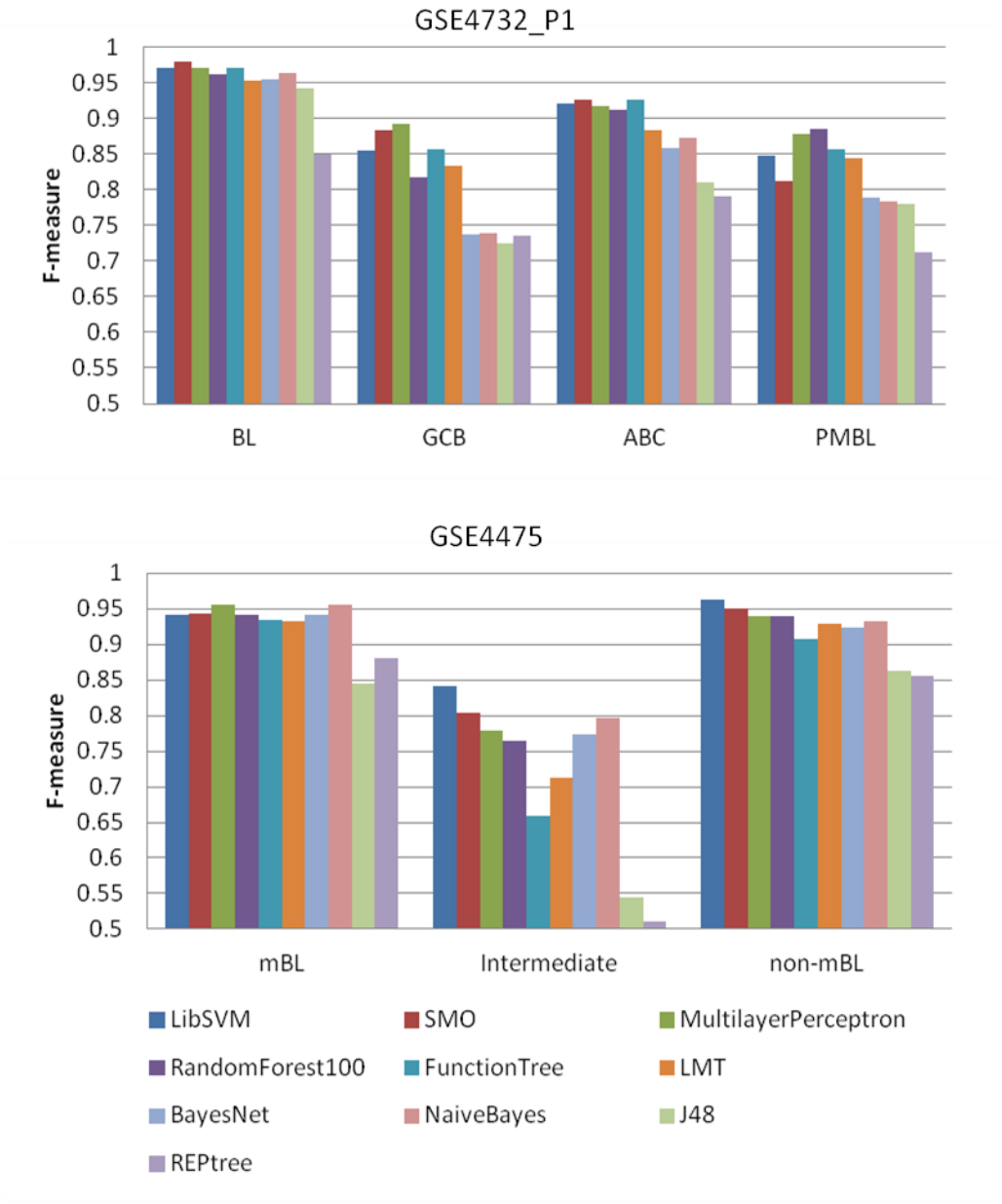


Figure 3-2: F-measure for each class of the classifiers built with a list of algorithms and tested on GSE4732_P1, GSE4475 multi-classification.

3.2.4. Comparison of algorithms in binary-classification

Next we assessed the efficiency of the algorithms in the binary classification situation, where all cases were assigned into BL or DLBCL, and this is also for the purpose of combining previous work and comparing between the two studies in the follow up analysis. In the Dave data all DLBCL subtypes were put into DLBCL class, so there are 54 BL samples and 249 DLBCLs. Two standards were employed in reassigning the 48 intermediate cases in the Hummel data: the ‘strict’ standard of BL definition, which put all intermediate cases together with the non-mBL cases that leads to 44 BL and 177 DLBCL; and the ‘wide’ standard assigns the cases by the BL probability generated by the Hummel classifier (probability greater than 0.5 BL, less than 0.5 DLBCL), which gives 59 BL and 162 DLBCL. The classifiers performance is shown in Figure 2-3.

In dataset GSE4732_P1, it is possible to achieve a very high DLBCL F-measure around 0.98 by all algorithms and an also high BL F-measure around 0.94, except J48 and REPtree algorithms. In dataset GSE4475 we investigated two definitions of BL: using the strict definition again very high DLBCL F-measure and BL F-measure are possible, while with the wider definition, the F-measure of DLBCL and BL drop down a little, indicating that the classes are less well defined in terms of gene expression when this standard is adopted. And the LibSVM algorithm showed clearly better performance in most situations.

The overall accuracy of the classifiers built under different conditions is listed in Table 3-2. And the average accuracy of the five conditions is compared among the algorithms. Given the level of the uncertainty in the actual classification of intermediate cases, we consider that these results reproduced the previous work at a level sufficient to support further investigations. Based on relative performance, LibSVM showed the best overall accuracy, and this is also consistent with the reports by many other groups that SVM usually gives the best performance. In addition LibSVM is a well-implemented library that can be easily used most working environments. Thus we chose support vector machines (SVMs) that implemented in LibSVM as our classifier method.

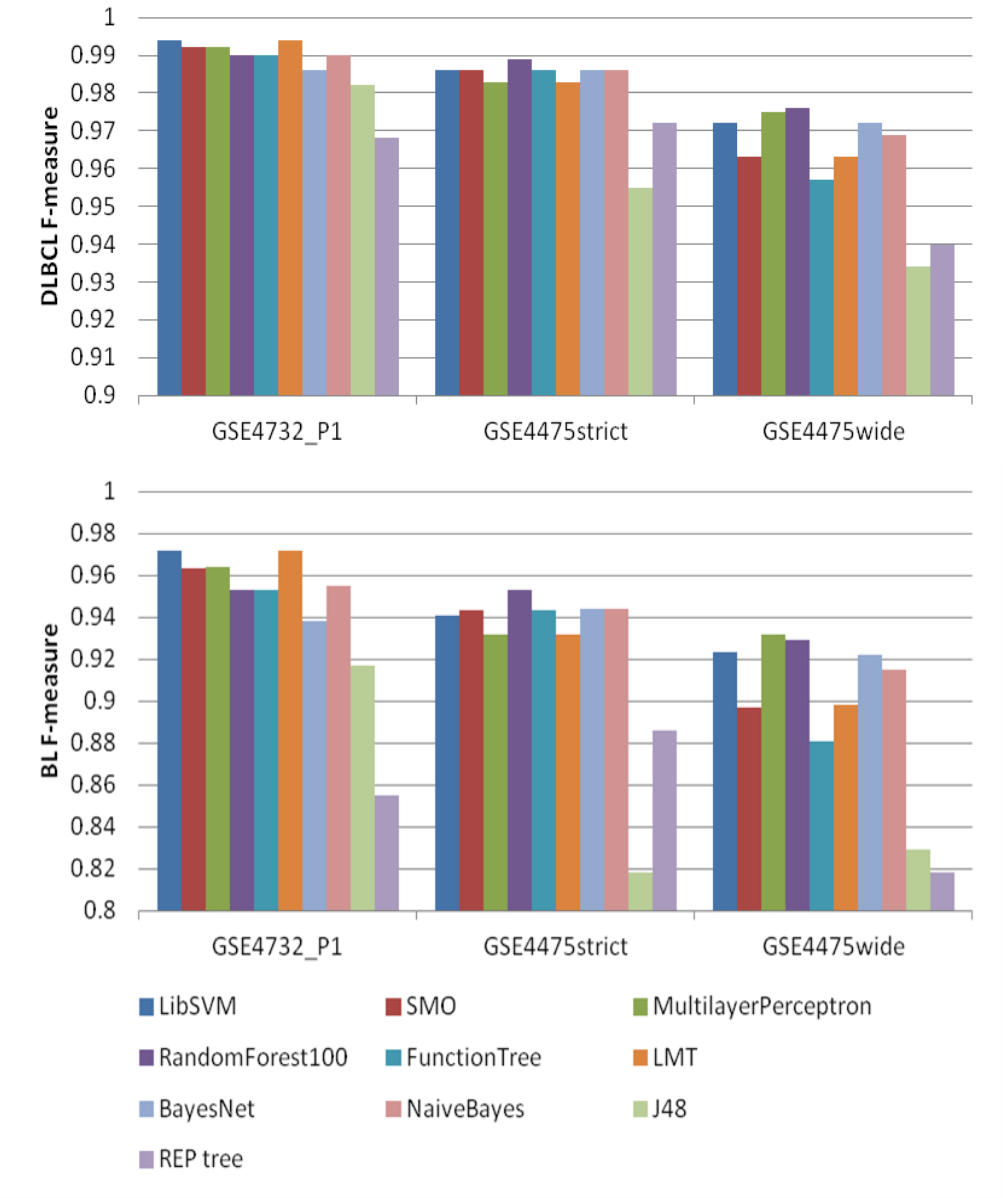


Figure 3-3: F-measure for each class of the classifiers built with a list of algorithms and tested on GSE4732_P1, GSE4475 binary-classification..

Table 3-2: Accuracy of 10-fold cross-validation for 10 algorithms in two data sets

Algorithm	GSE4732_P1: Multi-Class	GSE4475: Multi-Class	GSE4732_P1:Binary-Class	GSE4475: Strict	GSE4475: Wide	Average
LibSVM	0.8845	0.9186	0.9901	0.9819	0.9639	0.9478
SMO	0.8845	0.9277	0.9868	0.9774	0.9503	0.9453
MultilayerPerceptron	0.8911	0.896	0.9868	0.9729	0.9639	0.9421
RandomForest100	0.8383	0.9322	0.9835	0.9819	0.9639	0.9399
FT	0.8647	0.8824	0.9835	0.9774	0.9367	0.9289
LMT	0.8251	0.8688	0.9901	0.9729	0.9457	0.9205
BayesNet	0.7723	0.896	0.977	0.9774	0.9593	0.9164
NaiveBayes	0.7756	0.8824	0.9835	0.9774	0.9548	0.9147
J48	0.7327	0.8191	0.9703	0.9277	0.905	0.871
REP tree	0.723	0.8191	0.9472	0.9548	0.9096	0.8707

3.3 Choosing the optimal gene set

Gene (feature) selection is another important issue to consider in microarray data classification. How to choose differentially expressed genes and what is the final gene set that goes to the classifier can all make a difference in distinguishing classes. There are a wide variety of statistical techniques brought to this problem, ranging from simple t-tests, analysis of variance (ANOVA) to information gain, rank-based statistics, SAM and moderated t-statistics all give an excellent approach (details are introduced in feature selection section in chapter 2). However, different methods, particularly when applied in different datasets, may return differentially expressed genes that share limited overlap, and a reasonable way to find the genes that are biologically informative and best separate different phenotypes is to combine several methods and choose the genes selected by the majority. To address this concern, we compared the classifier genes selected by each group and derived a few gene sets that are potentially predictive. In addition a new list of gene signatures was identified as a complementary to previous work.

3.3.1. Find common genes

In the study performed by Dave group, the 217 genes used to build the classifier came from two parts. The first part consists of 21 *MYC* target genes identified by an OCI-Ly10 DLBCL cell line RNA interference experiment, and the second part consists of 100 genes that most significantly differentiate the 45 classic BL, which were originally diagnosed as BL and confirmed as such by the pathological review, from each DLBCL subtype by t-statistic. The Hummel group employed a core group extension algorithm by determining a shrinkage parameter of the nearest shrunken centroid feature selection method, and yield a gene set containing 58 genes (74 probes).

To investigate the genes used in each classifier and dataset, probes from Hummel dataset were annotated by the “hgu133a.db” [184] database in R. Gene symbols of the Dave dataset were annotated with an additional file (data4732.txt) provided by the author. Then all gene symbols were checked

with HGNChelper [185] R package and updated to the latest approved symbol if available.

Figure 3-4 and Table 3-3 both show the lists of the genes applied in classifiers as well as the genes being present and identifiable in the datasets. There were only 21 genes found to be in common between the two classifiers. Further, only 28 from 58 genes used by Hummel's classifier exist in Dave's data (GSE4732_P1), while 173 from 217 genes used in Dave's classifier appear in Hummel's data (GSE4475), and there were 1901 genes shared by both data sets.

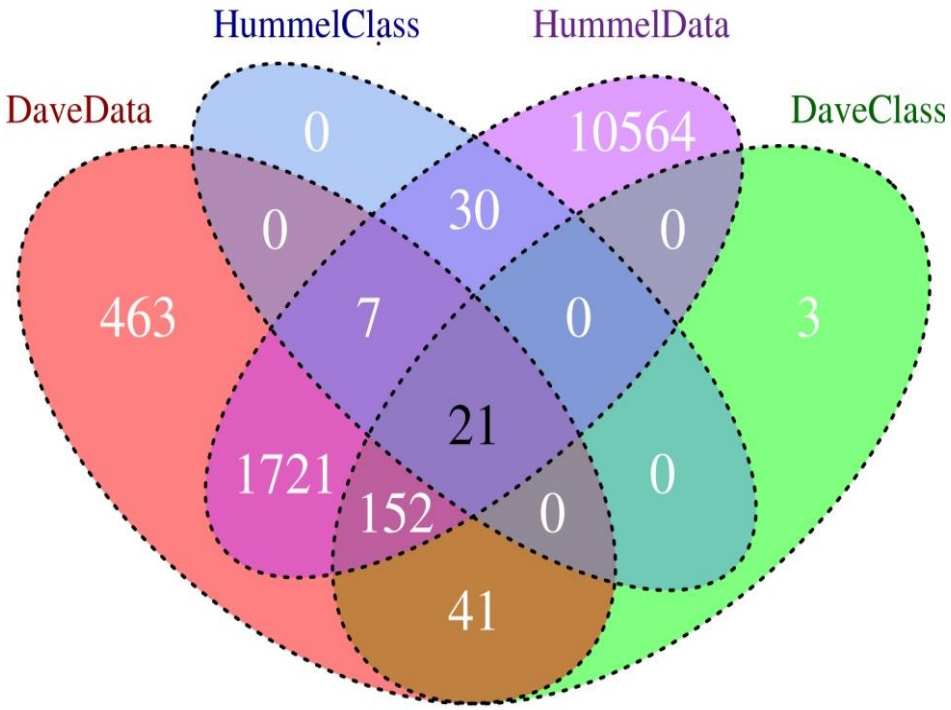


Figure 3-4: Venn diagram of genes derived from previous datasets and classifiers.

Table 3-3: Numbers of genes in data sets and used in classifiers

Genes from different places	GSE4732_P1	GSE4475	Overlap
HGNC matched genes on platform	2405	12495	1901
Genes used in authors' classifier	217	58	21
Classifier genes located in data[1]	214	58	21
Classifier genes available in other data set[2]	173	28	-

[1]. We were unable to locate all reported classifier genes in GSE4732_P1

[2]. Dave classifier genes available in GSE4475 and Hummel classifier genes in GSE4732_P1

3.3.2. Identify differentially expressed genes

We chose the 21-gene set, 28-gene set and 173-gene set derived from previous datasets to assess the performance of different gene sets used to build the classifier. In addition, we also tested a 10-gene set [48] used in a recent classifier that employs on FFPE data from the NanoString platform. The 10 genes were picked from the 58 genes applied in the Hummel classifier by the follow-up study in the same lab. However there were only 6 genes of the 10 genes found in GSE4732_P1 dataset, thus only 6 genes were adopted in the 10-gene classifier on this dataset.

To further exploit the gene set that best discriminate the two lymphomas, we identified a set of genes as a control set by performing a gene expression comparison between the high confidence BL and DLBCL in the two previous datasets, which are the consistent 44 BL with 235 DLBCL agreed by both pathological and molecular diagnosis in GSE4732_P1, and 44 mBLs plus 129 non-mBLs in GSE4475. In each dataset 200 most significantly differentially expressed probes were selected using limma package filtered with $\log_2$ fold change over 1. There are 56 genes that were picked in both datasets, but 4 genes that were used in both of the previous classifiers were not included. So a total 60 genes gene set was used as a comparison to

previously reported gene sets. The overlap situation of all five gene sets is illustrated in Figure 3-5 below.

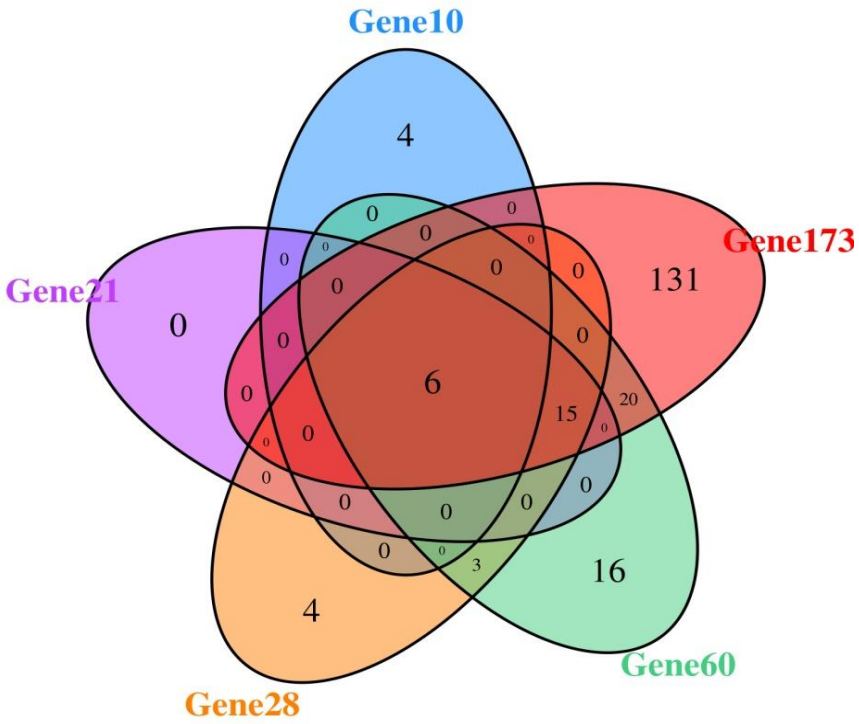


Figure 3-5: Venn diagram of five gene sets used to compare the performance of classifiers.

3.3.3. Comparison of different gene sets

Next different classifiers using the various gene sets were compared by 10-fold cross-validation of the two datasets GSE4732_P1 and GSE4475, using LibSVM algorithm with both default parameters and optimized parameters. There are three common kernels implemented in LibSVM: linear, sigmoid polynomial and RBF (radial basis function), and we chose RBF which is the default also most recommended kernel for below reasons: (1) it has fewer hyper parameters that influence the complexity of model selection, (2) it has fewer numerical difficulties and (3) other kernels behave like RBF at certain parameters. In the RBF kernel model, parameter optimization involves the

kernel parameter γ and the trade-off parameter C. We used the automatic script easy.py for a parameter grid search provided in the libSVM to select the optimized model parameters: the search range of C value was 2^{-5} to 2^{15} with a step of 2^{2}, the range of γ value is 2^{3} to 2^{-15} with a step of 2^{-2} and 5-fold cross-validation as assessment.

The performance of the classifiers built by each gene set with default and optimized parameters are showed in Figure 3-6. And it is evaluated with the F-measure of BL and DLBCL categories respectively. The figure shows that there is limited difference in the classifiers built by different gene sets, almost all of which can classify DLBCL and BL very well. And similarly to the test result among the algorithms, there is a better accuracy in dataset GSE4732 (DLBCL F-measure around 0.985 and BL F-measure around 0.94) and GSE4475 strict definition (DLBCL F-measure over 0.98 and BL F-measure over 0.92), with a relatively lower accuracy in GSE4732 wide definition (DLBCL F-measure around 0.96 and BL F-measure around 0.9). Comparing to the classifiers built with original gene sets (214 genes in the Dave classifier and 58 genes in the Hummel classifier), the other tested four gene sets classified the cases with a slightly decrease of accuracy depending on specific gene set.

Optimization of LibSVM parameters results in a modest increase of accuracy over the use of default parameters. In most cases the 28 gene sets match the performance of the full list in both data sets with only insignificant reductions in accuracy. The overall accuracy of the classifiers built by each gene set is listed in Table 3-4, and the correlation of the classification probability of all conditions are illustrated in Figure 3-7.

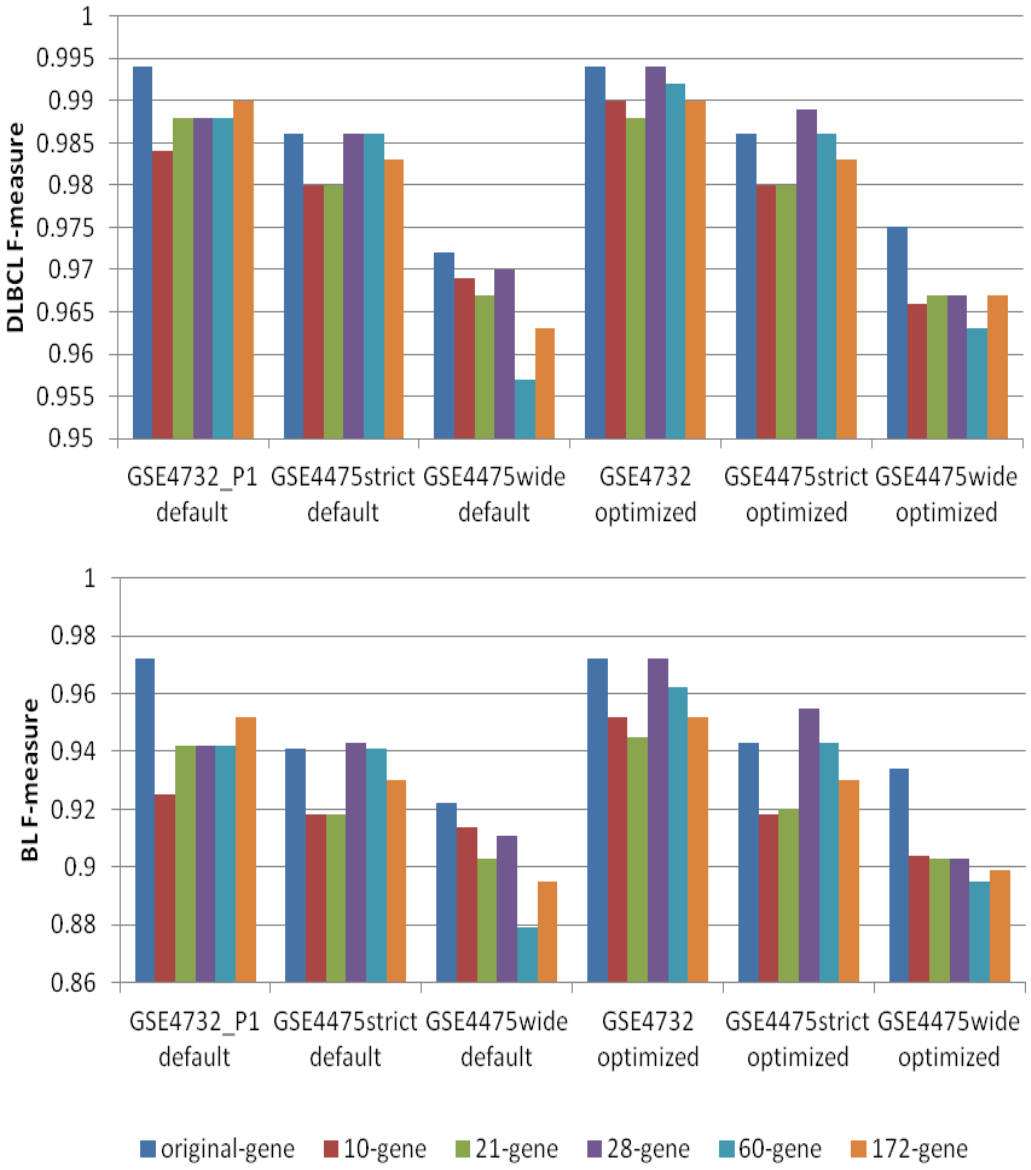


Figure 3-6: F-measure for each class of the classifiers built with a list of gene sets and tested on GSE4732_P1, GSE4475 binary-classification.

Table 3-4: Overall accuracy of tested gene sets in building the classifier

	GSE4732_P1	GSE4475strict	GSE4475wide	Average
10-gene	0.973	0.968	0.954	0.965
21-gene	0.976	0.968	0.95	0.965

28-gene	0.983	0.977	0.954	0.971
60-gene	0.98	0.977	0.936	0.964
173-gene	0.983	0.973	0.945	0.967

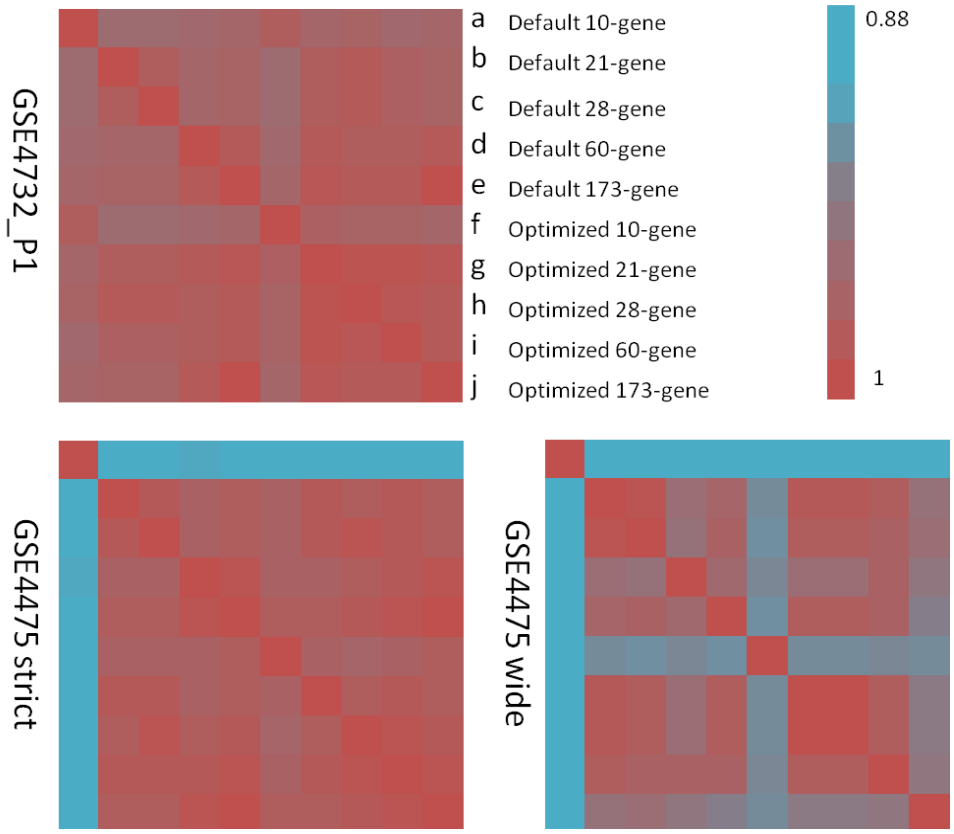


Figure 3-7: Classification probability correlation of the classifiers built with five gene sets and tested on GSE4732_P1 and GSE4475.

There is hardly any difference of the overall accuracy among the gene sets, and the classification probability based on all gene set either with default or optimized parameter are highly correlated, the least correlated is the 10-gene set however still have a correlation over 0.88. This shows that the gene set does not affect the classifier very much as long as it contains the basic informative genes and it is in a reasonable number. More importantly

they show conclusively that classifiers based on small gene sets perform at least as well as their larger counterparts. In particular there is no advantage to large gene sets, suggesting that shorter lists that are less prone to over fitting should be used. It seems the 28-gene set has the best average accuracy and a more stable performance according to Figure 3-6, and it was chosen for following analysis.

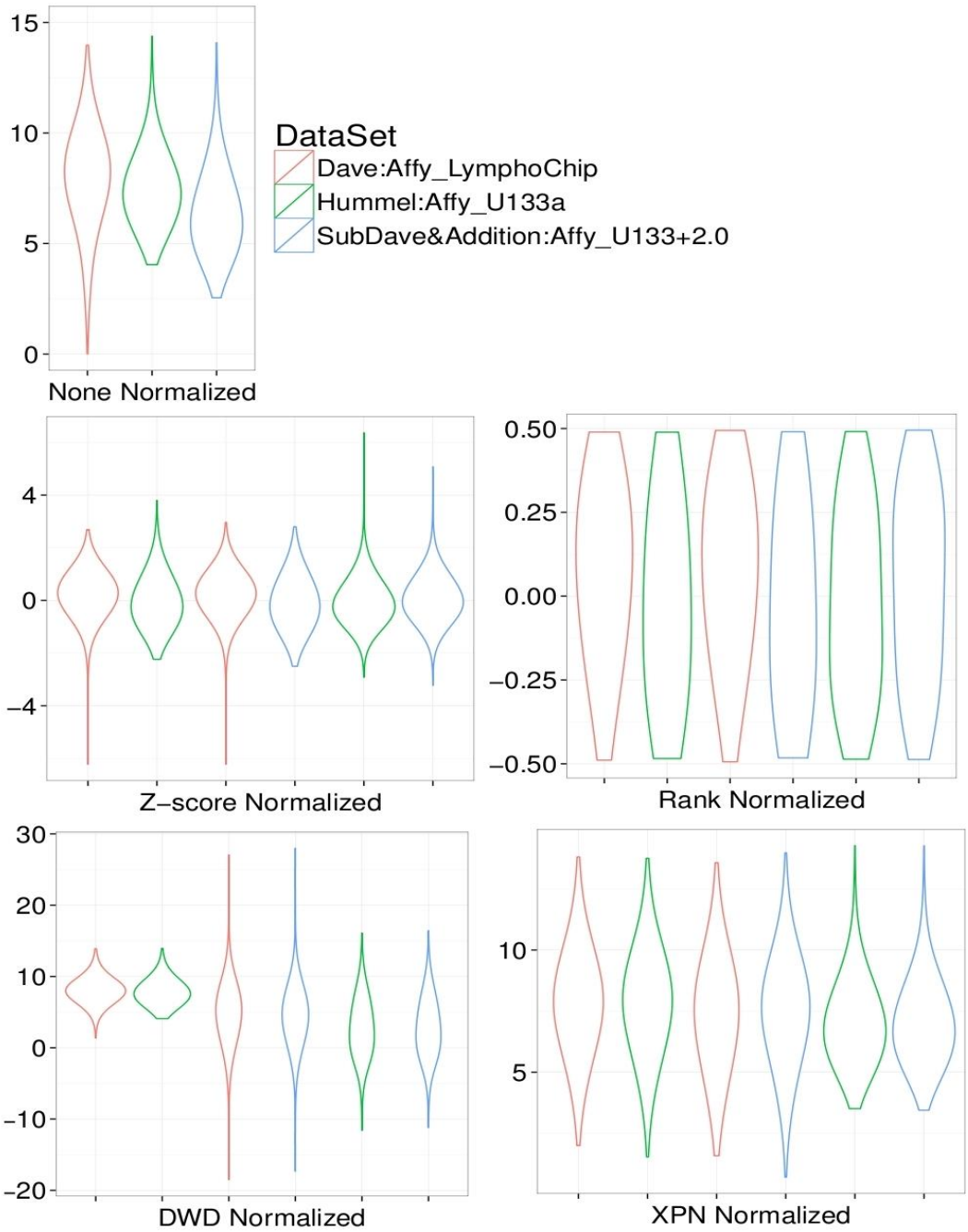
3.4 Cross-platform normalization and training set effect

Microarray platforms have developed and improved quickly over the years, although classifiers have been developed and validated in one single research. The variability of measurement among platforms leads to challenges for the combination and comparison between datasets. An effective solution is to perform cross-platform normalization so that data collected from various platforms can achieve a higher comparability. Several studies have provided methods related to this problem. Here we compared four cross-platform normalization methods: Z-score, rank-score and two more sophisticated methods XPN, DWD [145, 146].

3.4.1. Cross-platform normalization methods

Z-score normalization operates for each gene independently producing a normalised expression value for each sample as $\mathrm{z} = (\mathrm{x} - \mathrm{m})/\mathrm{s}$, where x is the un-normalised expression value of the gene and m and s are the mean and standard deviation of x over all samples. For rank-score normalization, $r = R/N - 0.5$ is the normalised value, where R is the rank of the sample with respect to the N other samples on the basis of the expression of the gene concerned. Z-score and rank-score normalisation have potential deficiencies, but also have the advantage of being applicable to data from methods such as RT-PCR and NanoString [186] which are designed to measure expression of only relatively small gene sets. XPN and DWD are two more complex methods implemented in the R CONOR package [147]; both are capable to give high consistency and low loss in gene detection according to the author, however require a relative large number of genes to perform a robust normalization.

Figure 3-8 shows the expression range of an example sample from each platform under four cross-platform normalization methods. Before cross-platform normalization, expression value from each platform has a distinct distribution and detection range, for example the main detection signal of Affymetrix U133a chip is higher than the Affymetrix U133+2.0 chip (Figure 3-8 None Normalized.) Although these four methods normalize the data in quite different ways, after normalization each pair of compared data sets gets a similar mean and distribution



3.4.2. Cross-platform normalization effect

The normalization methods were examined with the 28-gene set LibSVM classifiers. We trained the classifiers on either one of the previous two datasets GSE4732_P1 or GSE4475 with strict and wide BL definition both, and tested the classifiers on the other dataset as well as on additional four data sets: GSE4732_P2, GSE17102, plus two EBL and HIV-related lymphoma datasets GSE26673 and GSE17189. All of the four datasets were read and processed with the R “affy” package. The above 28 signature genes were extracted with expression values and merged with limma package. Table 3-5 shows the accuracy of testing the classifier using different data normalization methods.

The results of training and testing between GSE4732_P1 and GSE4475 showed that there is little difference among the above normalization methods, a classifier trained on GSE4732_P1 performs reasonably when tested on GSE4475 with the strict BL definition, giving error rates (specificity) around 9% for BL and <2% for DLBCL. Conversely, training on GSE4475 _strict and testing on GSE4732_P1 again gave good performance (errors around 4% for BL and 1% for DLBCL), indicating the classifier adopted on GSE4732_P1 corresponds to a BL criterion similar to the GSE4475 strict stratification. And as would be expected, training with the wide definition of BL in GSE4475 reduced the BL error rate observed when testing on GSE4732_p1 to 2% with a corresponding increase of the DLBCL error rate to around 5%.

Figure 3-8: Violin plot of the gene expression for an example case from each platform when normalized by different methods.

GSE4732_P2 is formed from a subset of the samples in GSE4732_P1 but from a different platform. It is surprising therefore that the classifier trained on GSE4732_P1 performed relatively poorly on this data set (BL error rates

15-21% depending on normalization method), and the classifier trained on GSE4475 performed worse (BL error rates of 27-33%). However a better classification result can be obtained by using a wider BL definition in the training set (error rate 3~6%), which suggest this could due to the reason that the two previous classifiers developed above adopt a narrower definition of BL, thus assigning cases with a weaker BL signal to the DLBCL category.

Table 3-5: Error rates for classifiers trained on one data set and tested on other public data sets

	BL error rate[1]				DLBCL error rate[1]			
Normalization	**Z-score**	**Rank**	**XPN**	**DWD**	**Z-score**	**Rank**	**XPN**	**DWD**
Train GSE4732_P1: Test on other data sets below								
GSE4475_strict[2]	0.09	0.09	0.09	0.09	0.017	0.017	0.006	0
GSE4732_p2	0.182	0.212	0.152	0.152	0	0	0	0
GSE10172_strict[2]	0.231	0.308	0.385	0.308	0	0	0	0
GSE26673 EBL	0.615	0.692	0.846	0.384				
GSE26673&GSE17189 HIV-related	0.833	1	1	0.667	0	0	0	0
Train GSE4475_strict BL definition: Test on other data sets below								
GSE4732_p1	0.04	0.04	0.04	0.04	0.012	0.008	0.012	0.012
GSE4732_p2	0.303	0.333	0.273	0.273	0	0	0	0
GSE10172_strict[2]	0.154	0.154	0.308	0.154	0	0	0	0

GSE26673 EBL	0.615	0.538	0.769	0.538				
GSE26673&GSE17189 HIV-related	0.833	0.833	1	0.833	0	0	0	0
Train GSE4475_wide BL definition: Test on other data sets below								
GSE4732_p1	0.02	0.02	0.02	0.02	0.04	0.05	0.06	0.07
GSE4732_p2	0.06	0.03	0.03	0.03	0.015	0.015	0.015	0.015
GSE10172_strict[2]	0.078	0.078	0	0.078	0.043	0.043	0	0.043
GSE26673 EBL	0.154	0.154	0.308	0.154				
GSE26673&GSE17189 HIV-related	0.5	0.333	0.833	0.5	0	0	0	0

[1]. Error rate is the specificity value for the indicated class

[2]. The sample in this dataset are assigned to mBL, intermediate, non-mBL three categories, here we set the strict BL definition as the standard which put intermediate and non-mBL together as DLBCL class

As to GSE10172, this is a smaller dataset generated by the same group who produced GSE4475, which were also classified into three categories, and here we used the similar strict definition as GSE4475 for the binary classes. Classifiers trained on either GSE4475 _strict or GSE4732_P1 produce zero error rate for DLBCL cases but higher errors for BL. Nevertheless, it is again that the classifier trained on the GSE4475 _wide produced a more accurate classification in GSE10172.

Figure 3-9 shows the prediction probability of GSE4732_P2 and GSE10172 when the classifiers were trained on GSE4732_P1, GSE4475_strict, GSE4475_wide definition respectively under four normalization methods. The left column is the actual class (assigned by authors) of each sample, with red color referring to BL, blue as DLBCL and green as intermediate class (for GSE10172 the column was separated into two, and the right half column indicates the probability from the original dataset). The other columns are the prediction results by different training set and normalization methods options, with red color representing high BL probability and blue color the opposite. This is again showed that that if the classifiers were trained by the same training set, the classification difference among the normalization methods is subtle. And that although the probabilities of the classifiers trained on GSE4732_P1 and GSE4475_strict are similar, they both predicted some of the BL cases as DLBCLs. However, the classification results trained on GSE4475_wide definition are more close to the original class.

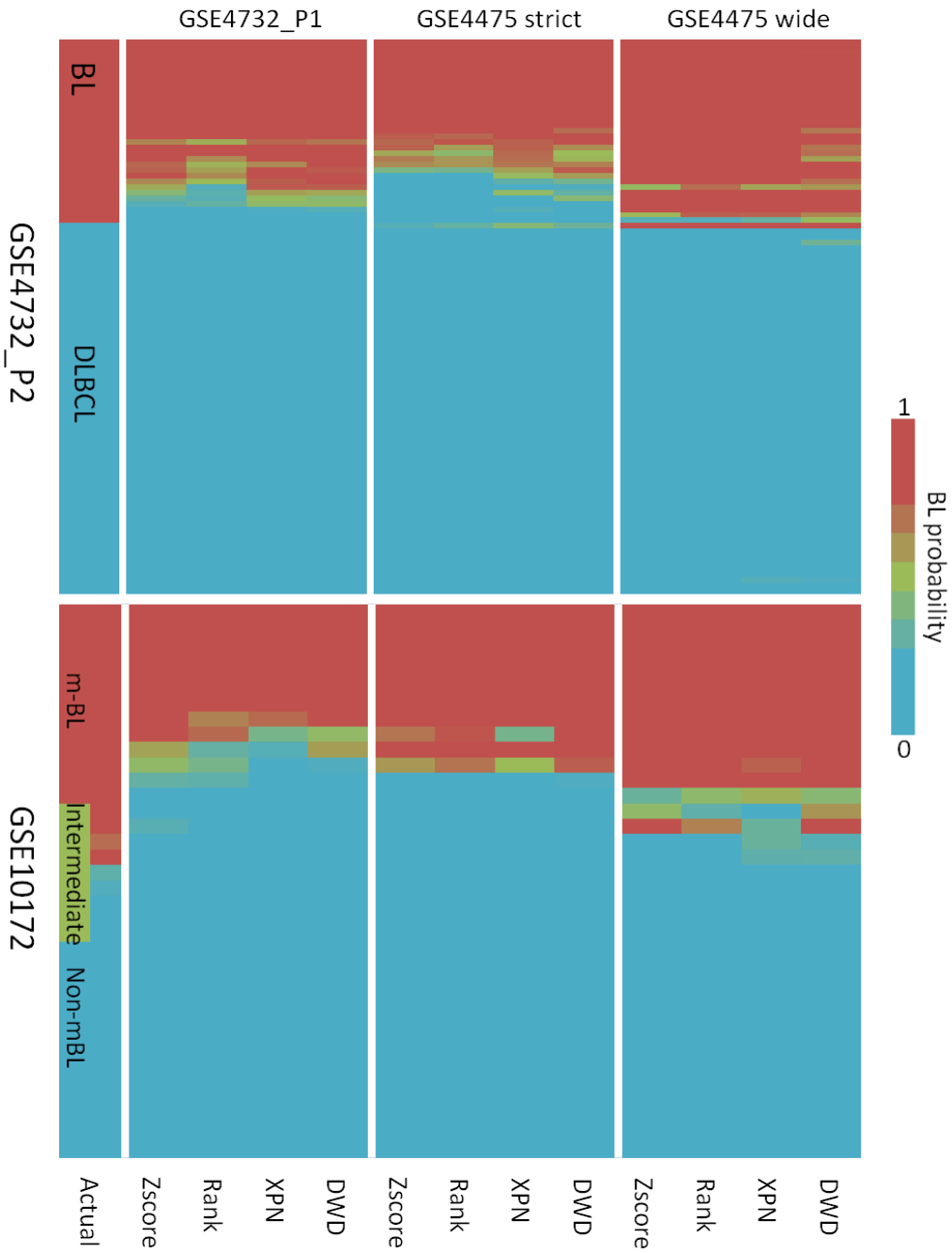


Figure 3-9: Prediction results overview of the classifiers when that datasets are normalized by a list of cross-platform normalization methods.

GSE17189 and 26673 contain EBL and HIV related BL cases in contrast to the sporadic cases from the other datasets. And once more the classifiers trained with strict definitions of BL perform poorly with this data (BL error rate > 50%). As GSE4475 strict definition only classified a case as BL when the BL probability is greater than 0.95 in the original classifier, and similarly the BL in Dave classifier are samples classified as BL by all predictors, it is not hard to understand the classifiers trained by the two definitions predicted some BL as DLBCL. However the classifier trained by GSE4475 wide gave results more close to the original classes in other datasets, which suggest the BL definition has a considerable effect on the classifiers.

3.4.3. Training set effects

The effect of BL definition in the training set on the classifier's performance is more thoroughly investigated in this section. We knew that classifiers trained on GSE4732_P1 and GSE4475_strict definition datasets perform badly in recognizing BL from other datasets. This could because that both datasets applied a rather strict definition of BL, and that a proportion of samples with less confidence of BL were assigned to DLBCL category. The original purpose is to keep BL class as clean as possible; however this may cause a bias when they are included in a training set.

In order to explore how much training set affect the classification result, we trained a list of classifiers with dataset GSE4475 by applying different thresholds according to the BL probability for each sample in the original Hummel's classifier. We set a range of thresholds to only include a subset of GSE4475 in the training set: a threshold of 0.9 means using only the BL/DLBCLs cases that have a BL/DLBCL probability greater than 0.9 are used to train the classifier. The thresholds were set at 0.95, 0.9 and 0.8 respectively, and the classifiers of each threshold were tested on the rest of five datasets, comparing with the classification result of GSE4475_strict and GSE4475_wide definition.

The classification probability of the classifiers trained with each threshold one dataset GSE4732_P1 is illustrated in Figure 3-10, and the Z-score normalized expression heatmap of the 28 classifier genes for the corresponding samples are showed below the probabilities. The Figure

shows that the strict definition trained classifier classify the sample into classes very close to the actual class, while the other four training thresholds gave similar BL probabilities will all classifying a small group of samples as DLBCL. In addition, from the gene expression heat map, we can see there is an area in the middle that the expression is neither confident BL nor clearly DLBCL, and the samples classified as BL by other thresholds exhibit a similar expression pattern as confident BLs, suggesting these are the intermediate cases have less confidence of DLBCLs.

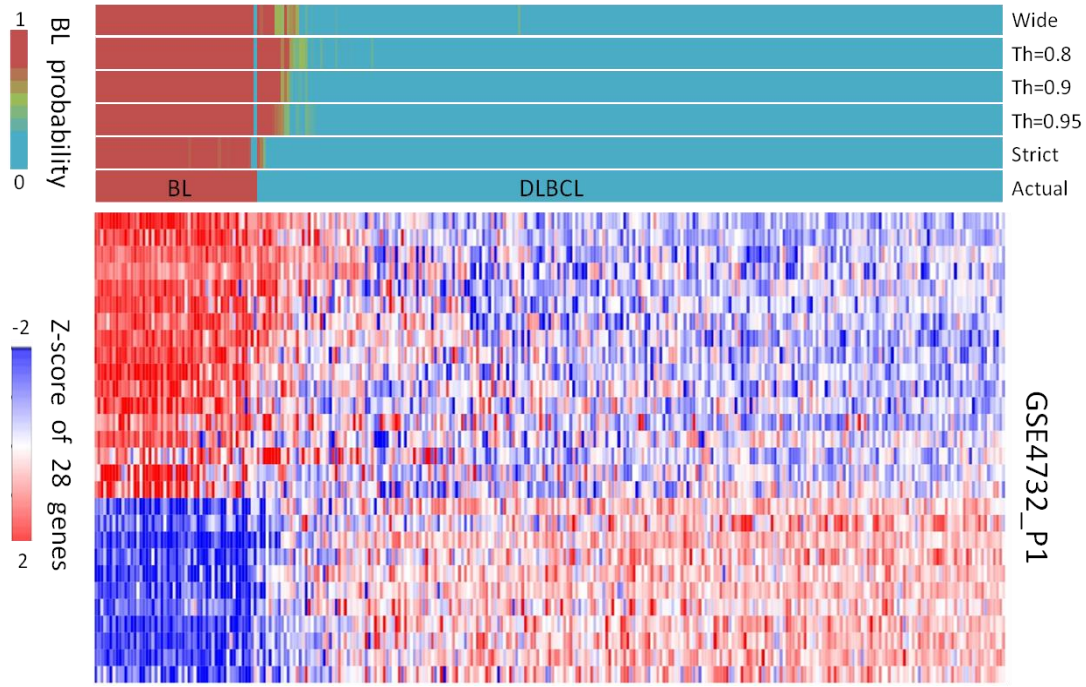


Figure 3-10: Performance of the classifier trained with different BL definition test on GSE4732_P1 with the heatmap of Z-score normalized 28 classifier genes expression value.

The classification probability results by different training set thresholds of the other four datasets: GSE4732_P2, GSE10172 plus EBL and HIV related BL/DLBCL datasets GSE17189 and GS26673, as well as their classifier gene expression heatmaps are illustrated in Figure 3-11. In GSE4732_P2

and GSE10172, it is the same situation that the strict defined training set classifies less cases as BL than other four thresholds, and these thresholds shows similar BL probabilities on each sample. And it is even more obvious that the cases called DLBCL by strict training set but BL by other are more likely to be BLs according to the expression heatmap.

Figure 3-11 for GSE17189 and GSE26673 also shows that EBL cases have a similar gene expression pattern to the sporadic cases but generally with a weaker signal, explaining the high error rates from the strictly trained classifiers and the improvement in this when a wider definition is applied. Many HIV related BL cases on the other hand appear to have gene expression patterns related at least as strongly to DLBCL cases as they are to sporadic BLs and do not classify as BL with any choice of training data. Although sharing many pathologic features with sporadic Burkitt lymphoma, the endemic and HIV associated Burkitt lymphoma cases do have a distinct pathogenesis and gene expression. Some classifiers can recognize EBL seemingly well, but we suggest that training these classifiers on data for sporadic BL and applying it to endemic or HIV related BL would not be advised. Given the distinct clinical settings of these disease variants this does not pose a significant issue in relation to development of an applied gene expression based classification tool.

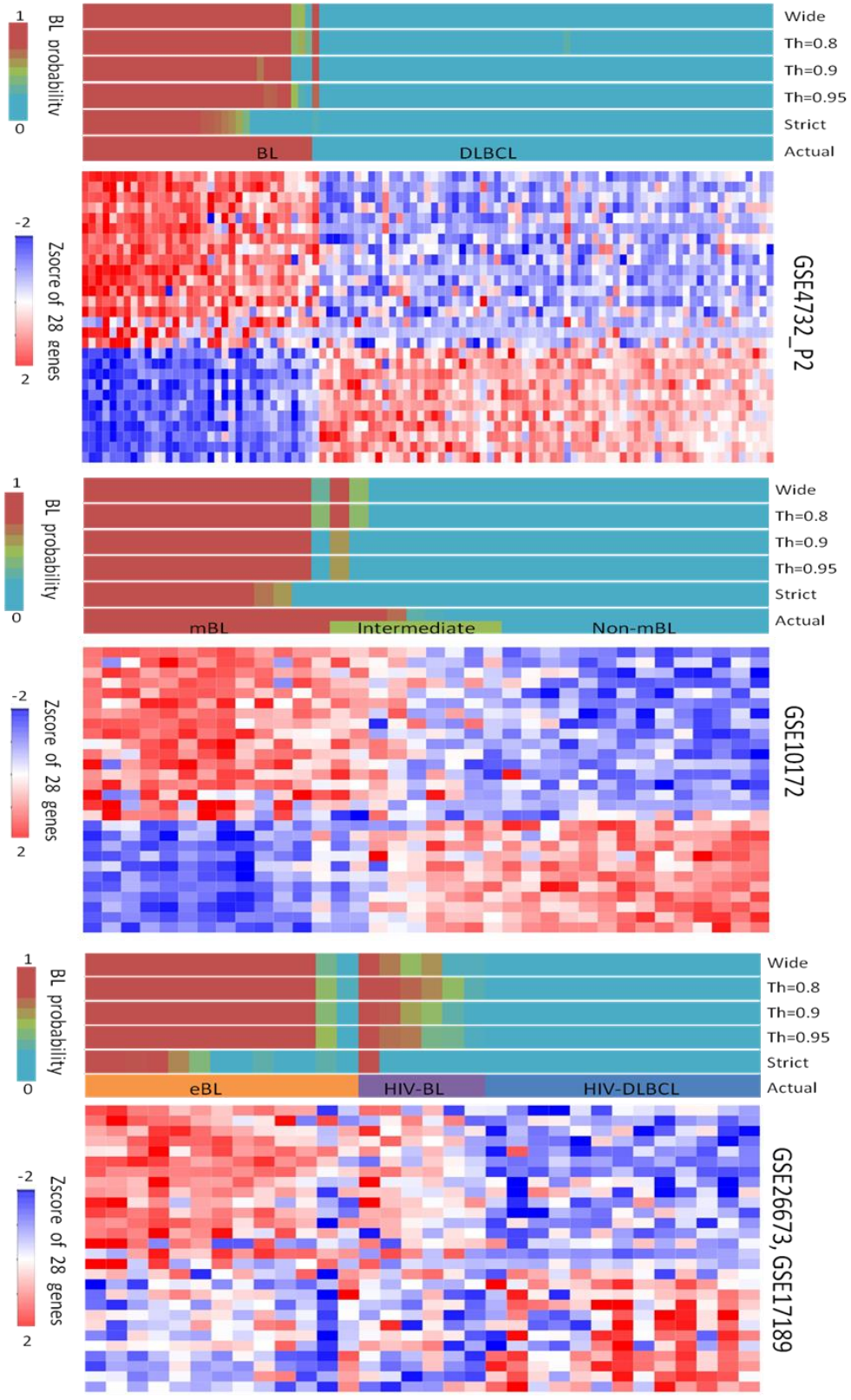


Figure 3-11: Classification results of GSE4732_P2, GSE10172, GSE17189 and GSE26673 when the classifier was trained with different BL definition plus the heatmap of Z-score normalized 28 classifier genes expression value.

Table 2-5 summarizes the classification accuracy of all five test datasets tested on above training set thresholds. The result shows that when tested on GSE4732_P1, other training set have a lower accuracy compared to strict definition. However on the rest of the datasets GSE4732_P2, GSE 10172, GSE26673 and GSE17189, the other training thresholds gave a better accuracy than the strict definition. There is no obvious difference on the training set thresholds apart from the strict definition; however the classifier achieved the best average accuracy when set the threshold at 0.95. And it also seems from Figure 3-11 that 0.95 threshold gave the classification BL probability most close to original dataset.

Table 3-6: Accuracy of the classifiers trained by a range of training thresholds.

Dataset	Strict	0.95 threshold	0.9 threshold	0.8 threshold	0.5 threshold
GSE4732_P1	0.987	0.967	0.96	0.96	0.954
GSE4732_P2	0.899	0.96	0.96	0.97	0.97
GSE10172	0.889	0.972	0.972	0.944	0.944
GSE26673	0.4	0.933	0.867	0.867	0.867
GSE17189	0.765	0.824	0.824	0.882	0.824
Average	0.788	0.9312	0.9166	0.9246	0.9118

3.5 Conclusion

By comparing a list of classification algorithms, various gene sets, different normalization methods, and cross-validation on one dataset or training on one dataset then test on others, we have presented a comprehensive study in establishing a robust GEP classifier. To conclude, these studies show two substantially different classifiers in the research literature have a high degree of concordance and that their results can be recapitulated, at least within the level of uncertainty associated with intermediate cases. Also that a classifier that uses fewer genes can effectively recapitulate the results from previous

classifiers, which is a substantial advantage making the classifier less prone to over-fitting.

The studies also show that the classifier can be successfully applied to other public datasets when cross-platform normalization methods are adopted, and the choice of the methods doesn't have much effect on the performance of the classifier. However the classification for cases with a less strong signal of BL is critically dependent on where the boundary between classes is placed in a spectrum of similar cases in the training set. In this study we investigated of the effect of training set in detail, and the final decision was to train the classifier on the two-way definition of BL based on the original class of GSE4475 (which is to set the threshold at 0.95).

This type of exploration on training set effect is noteworthy, since classifiers developed so far have largely been trained and tested within single data sets. And the dependence on training data highlights the underlying difficulty in this and many similar studies, which is the lack of a 'gold standard' against which to evaluate new classifiers.

Chapter 4
Validation of the classifier on in-house data from FFPE specimens

Despite the advantages of GEP classifier in distinguishing Burkitt lymphoma and diffuse large B-cell lymphoma, it has not yet been commonly used in routine clinical practice. The current diagnosis usually combines features such as morphology, phenotype and genetic aberrations identified by the fluorescence *in situ* hybridization (FISH) method based on the World Health Organization (WHO) guidelines, and no single parameter is regarded as a gold standard. One key reason for this is that until now almost all studies used fresh-frozen samples in research laboratories while clinical archive samples are mostly formalin-fixed paraffin-embedded (FFPE) tissues. FFPE data is likely to remain the clinical reality in the future, particularly in diagnostic laboratories responsible for large geographical areas with many hospitals. However the RNA extracted from these samples is normally degraded which limits the amount of biological information that can be derived, and makes gene expression measurement more difficult and error prone.

Recent advances in methods for extraction and assay of RNA from FFPE samples have achieved an encouraging reliability and reproducibility. Illumina technology developed a WG-DASL (Whole Genome DNA-mediated Annealing, Selection, extension, and Ligation) assay method, which combines the unique PCR and labelling steps of the original DASL assay with gene-based hybridization of Illumina's whole-genome probe set Direct Hybridization assay (see Figure 4-1). The DASL assay uses two short sequences labelling the target genes that can be separated by an arbitrary gap, which enables more flexibility in choosing the genes. In addition, DASL arrays use probes about only 50 bases instead of long sequences used in other cDNA arrays that require good quality of RNA [187]. This unique methodology greatly increases the DASL assay target set, while retaining

the ability to accurately profile low-abundance and partially-degraded human RNA samples, such as RNAs derived from FFPE tissues.

Figure 4-1: Illumina Whole-Genome DASL HT Assay combines original DASL Assay and Direct Hybridization Assay.

Here we collected over a thousand FFPE samples from general clinical practice and other clinical trials in cooperation with the Haematological Malignancy Diagnostic Service (HMDS) department of St. James University Hospital. The gene expression profiles of the samples were measured by Illumina Whole-Genome DASL Assays. All samples were diagnosed as BL or DLBCL beforehand according to the currently applied diagnostic criteria. The datasets were used to validate the classifier developed in the last chapter, and to examine the concordance between the molecular prediction and current clinical diagnosis. Furthermore, where of the samples have detailed clinical information (genetic alteration, treatment management and survival time) available, we explored the survival indication of the molecular classifier as well as its potential clinical benefits.

4.1 FFPE data and preprocessing

4.1.1 . Data sets summary

Two series of experiments were performed by the WG-DASL Assay (based on the Illumina Human Ref-8 Version_3 BeadChip) and the WG-DASL HT Assay (based on the Illumina Human HT-12 Version_4 BeadChip) respectively. The Version_3 BeadChip is an earlier version of whole genome array that offers parallel analysis for 8 samples at once, and Version_4 BeadChip is the latest chip with throughput data up to 12 samples on each BeadChip. The Version_4 chip covers more well-characterized genes, gene candidates, and splice variants. The detailed probes contained in both chips are illustrated in Table 4-1. Also the Version_4 chip is uses improved reagents and protocols, and yields high self-reproducibility even with lower RNA input [188]. Although a high concordance between the performance of Version_4 chip and Version_3 chip was reported in the assessment of 16 commercially obtained FFPE samples[187], evaluations from such small numbers may not reflect the actual situation in large datasets.

Table 4-1: Detail probes comparison between Illumina BeadChips

Probes	Version3	Addition	Version4
Coding transcripts, well established annotations	23811	3442	27523
Coding transcripts, provisional annotations	426		426
Non-coding transcripts, well established annotations	283	1317	1580
Non-coding transcripts, provisional annotations	26		26
Total	24526	4759	29285

In our study, there are 456 samples performed by the WG-DASL Assay here is referred as Version_3 dataset, and 932 samples performed by the WG-

DASL HT Assay here is referred as Version_4 dataset. Both datasets contain a certain number of samples from the same patient repeated on one platform that allow us to examine the reproducibility of the particular chip, and there are also samples repeated on both platforms so that we could compare the consistency between Version_3 and Version_4 chips.

4.1.2 . Control probes

As discussed in chapter 2, quality checks of the data obtained from a specific array before further analysis is absolutely crucial, because even a single or few abnormal arrays can completely compromise the results of the analysis of a large data set. Arrays that do not meet the necessary quality standards should be discarded from the following analysis. This is even more of the case for FFPE samples, as the RNAs of which are usually less qualified to extract meaningful biological information.

Both of the Version_3 and Version_4 Illumina BeadChips contain a set of internal control probes to evaluate the quality of a single array. Poor performance measured by sample-dependent control probes could indicate a problem related to sample quality or labelling, and failure of the standard for sample-independent control probes may indicate a general problem with the hybridization, washing or staining. The control probes on each chip are listed in Table 4-2.

Table 4-2: Number and types of control probes on Version_3 and Version_4 s

Type	Description	Version_3	Version_4
Housekeeping	Sample quality controls	16	16
Negative	Background noise controls	309	290
Annealing	Labelling controls	10	10
Cy3-hyb	Hybridization controls	6	6
Low-stringency	Stringency controls	8	8
ERCC	Spike-in mix controls	0	92

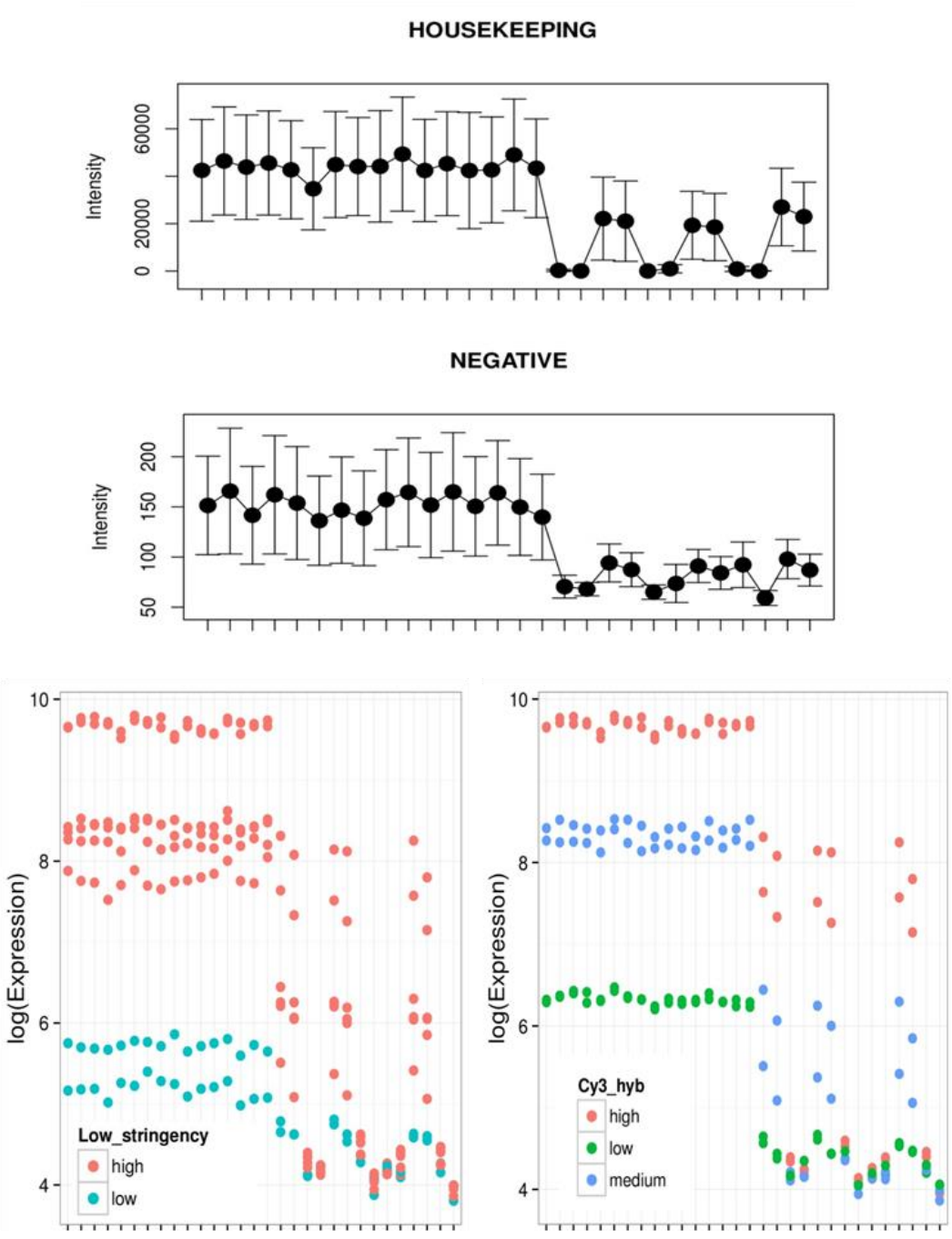


Figure 4-2: Examples of controls probes from Illumina WG-DASL chip.

The housekeeping probes should be expressed at a relatively high level in all samples, and negative controls are random sequences not used as target genes, which should be expected to have a low level signal. Cy3-hyb and low-stringency probes are controls for hybridization that present at different concentration levels, and the high level probes should have higher signal than low level probes. Annealing and spike-in probes are used to assess the sample labelling, which should be consistent in most samples. An example of the control probes in good and bad qualities is illustrated in Figure 4-2; the successful arrays show the corresponding characters for the controls mentioned above, while arrays with poor quality tend to have significant lower intensities for all conditions.

4.1.3 . Quality check

The quality of an array can be assessed by the relative comparisons with other arrays using the control probes. Here we use lumi package [144] in R to visualize the control probes and evaluate the arrays quality. The common types of control probes in both Illumina BeadChips were used as metrics to perform the quality check, and also we checked the detected number of probes on each array to exclude those don't have enough information. In addition, the arrays of poor quality were detected with the *detectOutlier* function in the lumi package. The outlier detection function is based on the distance from the sample to the centre (average of all samples after removing 10 percent samples farthest away from the centre).

For metrics of detected probe numbers and different control probes, if an array falls out of 1.645 × standard deviation from the mean (which will include 90 percent of all samples), it is considered to fail this specific metric. And the array is regarded as an outlier if it is picked out by *detectOutlier* function, or it fails probe detection or housekeeping detection or either of the other two types of the control probes. Figure 4-3 shows an example of 30 samples from Version_3 chip on each quality check metric, and the samples with red colour are those failed the corresponding metric.

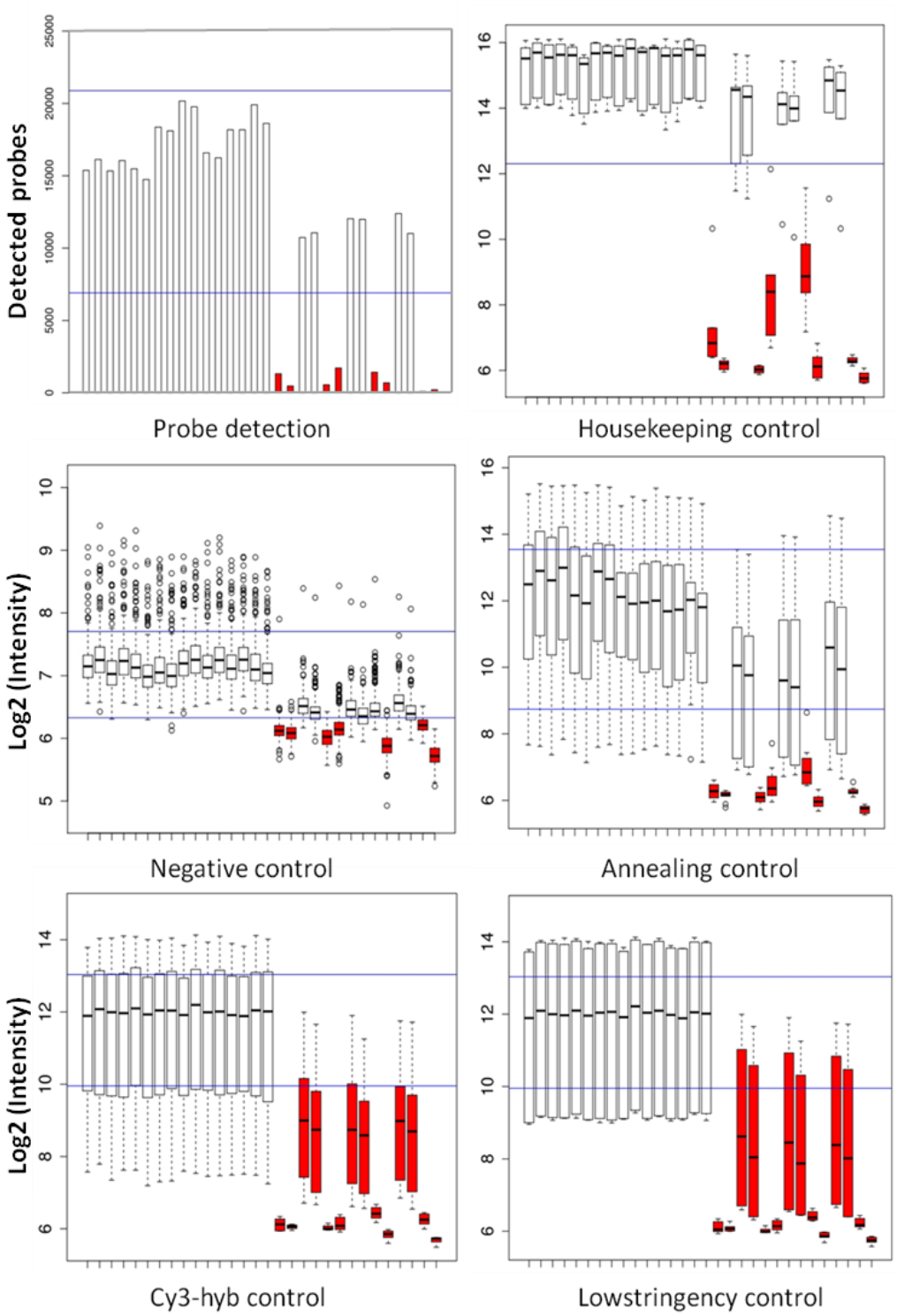


Figure 4-3: Quality check on each metric of 30 samples from Version_3 chip.

The figure shows that most of the arrays that failed on each metric are usually the same ones, which indicates this is often the problem caused by the sample itself. A sample that has very low detected intensity or low housekeeping probe intensities tends to fail other metrics too. So the quality check step is able to catch out arrays that are not qualified for further analysis. As a result, 97 arrays and 119 arrays were removed from Version_3 and Version_4 dataset respectively. Table 3 lists the detailed number of arrays failed on each quality check.

Table 4-3 : Summary of outliers of each metric

Metric	Version_3	Version_4
Detected probe number outliers	52	52
DetectOutlier function outliers	83	84
Housekeeping control probe outliers	19	36
Negative control probe outliers	37	53
Annealing control probe outliers	40	47
Low-stringency control probe outliers	16	27
Cy3_hyb control probe outliers	16	13
Final outliers exclude from further analysis	97	119

4.1.4 . Preprocessing

After removing poor quality arrays, the raw intensities of the successful samples need to go through preprocessing steps before interpreting biological meanings, which normally includes background correction, normalization and summarization of expression values (detail discussed in chapter 2). The preprocessing of the Illumina DASL Version_3 and Version_4 data were carried out using lumi package in this study.

The raw intensities were adjusted with background information by *lumiB* function bg.adjust method, which estimates the background based on the

negative control probe information. Data was then transformed with the VST method [140] by the *lumiT* function: this is a critical step for Illumina data for it takes advantages of larger number of technical replicates available on the array. Then the samples were normalized with quantile normalization methods by *lumiN* function, and the probe expression was extracted by the *exprs* function. In the situation where several probes represent a single gene, the expression of this gene was summarised by *averep* function in limma [189] package.

4.2 Validation of the classifier on the Version_3 and Version_4 datasets

The Version_3 and Version_4 datasets were used to test the classifier developed from fresh-frozen tissues of its performance on FFPE samples. In chapter 3 we examined the gene sets and classification algorithms to build the classifier, as well as the effect of cross-platform normalization method and the BL definition in the training set on the classification result. Here in this chapter we validated the 28-gene set and LibSVM the classifier which were chosen the as best options from previous studies. To further investigate the transferability of the classifier on FFPE samples, we compared the effects of different normalization methods and training set options on the Version_3 and Version_4 as well.

4.2.1. Classification Result of Version_3 dataset

The Version_3 dataset contains 359 samples of expression profiles for 18301 unique genes. Classifiers were trained and compared on four datasets GSE4732_P1, GSE4475_strict, GSE4475_wide and GSE4475_0.95th (training threshold equals 0.95, which only the cases have the classification confidence over 0.95 from GSE4475 are included in the training set, which are the original mBL and non-mBL cases). For each training set option, the samples and Version_3 test dataset were normalized by four cross-platform normalization methods: Z-score, Rank-score, XPN and DWD. The correlation of the classification BL probability from total 16 conditions is illustrated in Figure 4-4.

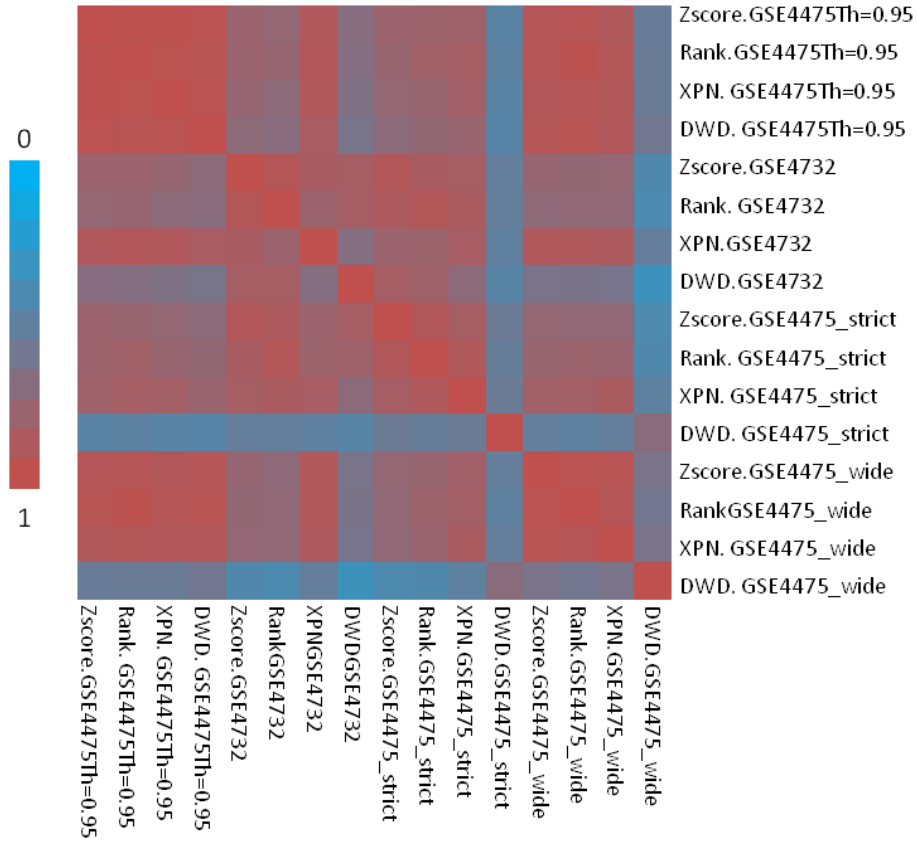


Figure 4-4: Correlation of Version_3 data classification BL probability by different normalization methods and training set options.

Figure 4-4 shows that DWD method normalized data somehow give low correlations with other methods, especially when the classifier is trained with GSE4475_strict and GSE4475_wide (correlation with other methods is around 0.4~0.6). This may suggest DWD normalization method works less effectively on Illumina DASL performed FFPE samples. It also shows that the classifiers trained with GSE4732_P1 and GSE4475_strict are with high correlation, and that GSE4475_wide and GSE4475_0.95th trained classifier are highly correlated.

However similarly to the test result on public datasets, the classifiers trained by GSE4732_P1 or GSE4475_strict predict fewer cases as BL, comparing with the classifiers trained by GSE4475_wide or GSE4475_0.95th. We also test the training threshold (explained in section 3.4.3) of 0.9 and 0.8 on

GSE4475 dataset, and the number of predicted BL and DLBCL is listed in Table 4-4. GSE4732_P1 and GSE4475_strict trained classifier only recognised less than 25 BLs, while other training sets recognised about 60, which indicates that as expected the more strict definition of BL in the training set leads to a narrower definition of BL of the classifier.

4.2.2. Classification Result of Version_4 dataset

There are 813 samples that have the expression 20818 unique genes in the Version_4 dataset, and the validation of the 28-gene set LibSVM classifier was done the same way on Version_3 dataset for the normalization methods and training set options. The correlation of the classification BL probability from different test conditions is illustrated in Figure 4-5.

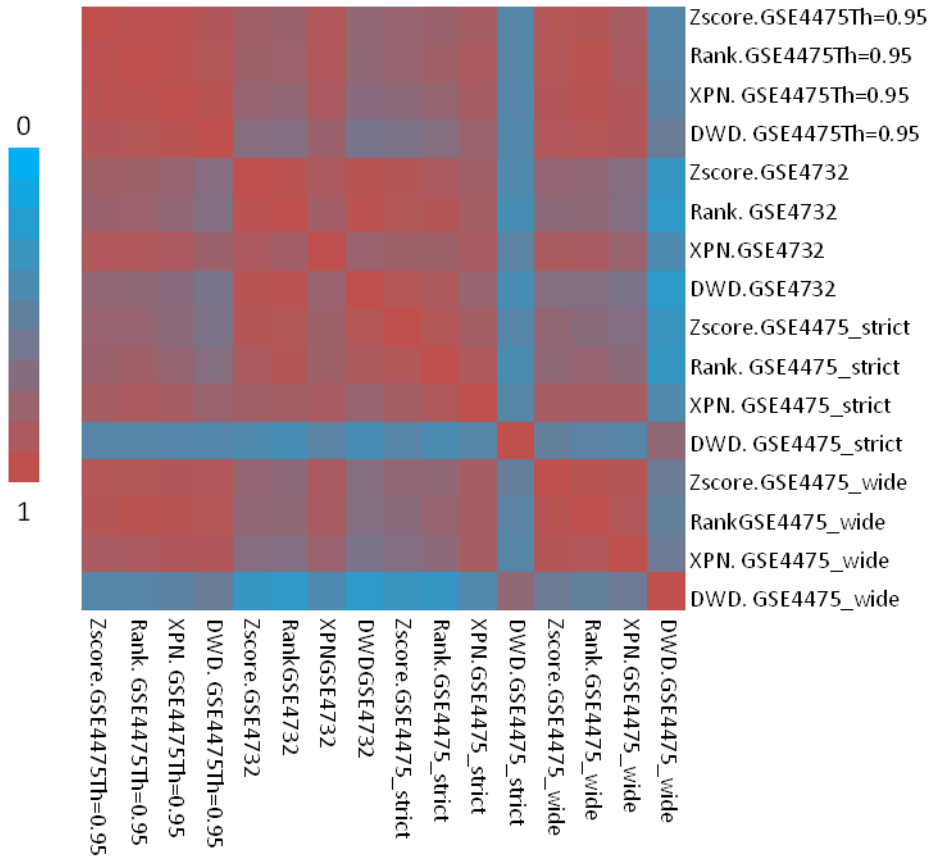


Figure 4-5: Correlation of Version_4 data classification BL probability by different normalization methods and training set options.

The Version_4 dataset test results again show that the classification BL probability of the data normalized by DWD method is less correlated with that by other methods. Nevertheless, the predicted class is strongly depending on the BL definition in the training set (see Table 4-4). From the studies on the public datasets in chapter 3, we know that the stricter definition of BL in GSE4732_P1 and GSE4475_strict would classify the samples with weaker signal of BL as DLBCL. As to FFPE data, it is more likely the BL samples express weaker signal than fresh-frozen tissues, because by its nature BL is prone to rapid degradation, hence the classifiers trained with strict definition predict those as DLBCLs.

Table 4-4: Number of predicted BL and DLBCL in classifiers trained with different training options

	Version_3		**Version_4**	
	BL	**DLBCL**	**BL**	**DLBCL**
GSE4732_P1	25	334	62	751
GSE4475_strict	23	336	48	765
GSE4475_wide	62	297	129	685
GSE4475Th=0.95	60	299	111	702
GSE4475Th=0.9	63	296	128	686
GSE4475Th=0.8	70	289	168	646

Figure 4-6 presents the Z-score normalized Version_4 data expression of the 28 classifier genes, as well as the classification BL probability by different training set options. The expression is more noisy compared with that of the fresh-frozen samples, however it shows a similar gradient from BL to DLBCL with no clear separation between two classes, and the cases that classified as BL by GSE4475_0.95th but as DLBCL by GSE4732_P1 and GSE4475_strict training sets exhibit an expression pattern very close to the confident BLs agreed by all classifies. Although there is no comparison as yet with any different diagnosis method, we use the classification result by

the GSE4475_0.95th, for it looks like this is more sensitive to the BLs with weaker signal from the FFPE samples according to the expression heatmap showed in Figure 4-6.

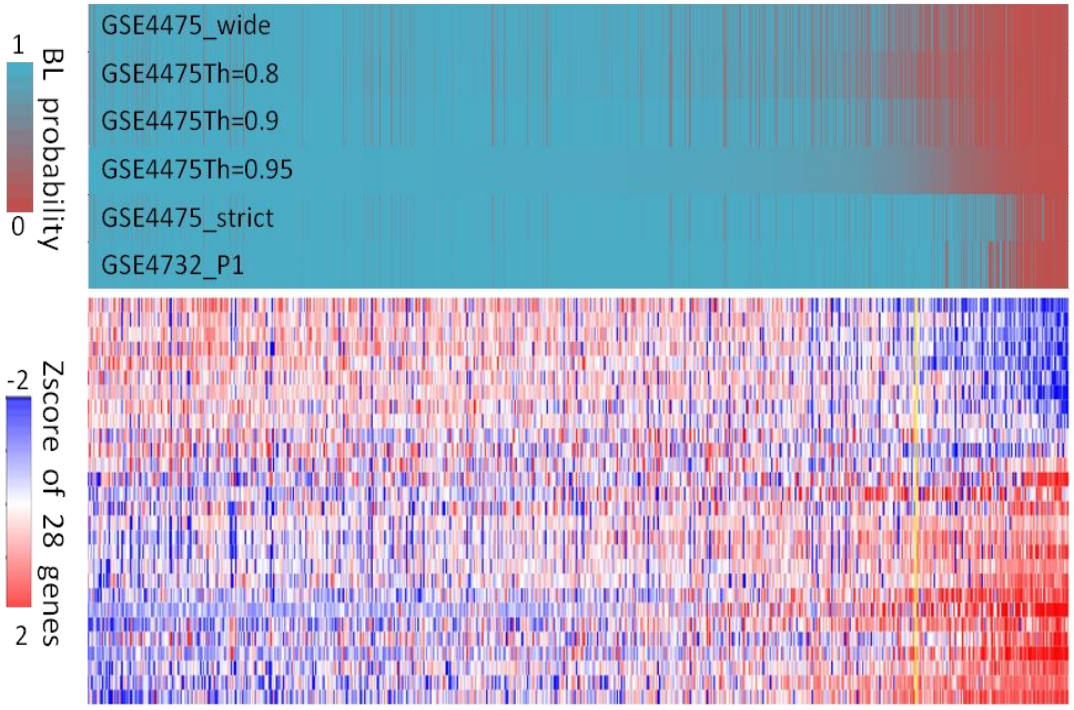


Figure 4-6: Classification results of Version_4 data when the classifiers are trained with different BL definition plus the heatmap of Z-score normalized 28 classifier genes expression value.

4.2.3. Reproducibility of Version_3 and Version_4 data

Following above investigations, we examined the reproducibility of the samples from both Illumina DASL chips using the results of the 28-gene LibSVM classifier trained by GSE4475_0.95th data. Of the 1172 samples in the Version_3 and Version_4 datasets together, 1083 (92.4%) have the same classification results (predicted classes) independent of normalization methods. For some cases the datasets contain replicates: in Version_3 there are 60 cases, in Version_4 69 cases, and in total 184 cases have replicates on either version (including some cases that are not replicated within a version, but have been done on both versions). Table 4-5 lists the

average BL probability variance of the replicates on each version and two versions together, when the data are normalized in different methods. It shows that the Z-score normalization produces the smallest variance, and this was used subsequently.

Table 4-5: BL probability variance of replicates of different normalization methods

	Z-score	**Rank**	**XPN**	**DWD**
Version_3 replicates average variance[1]	0.009	0.012	0.016	0.017
Version_4 replicates average variance	0.007	0.008	0.010	0.014
Version_3 Version_4 together average variance	0.008	0.010	0.014	0.012

[1]Average variance equals to the sum of BL probability variance on the replicates for each patient divided by total patient number.

There are 81 cases that were performed on both versions, and 3 (3.7%) have a different classification, 5 (6.2%) have a BL probability variance between the versions greater than 0.2 according to the Z-score normalized GSE4475_0.95th trained classifier. The detailed results for all 184 replicated cases are shown in Figure 4-7 (lower), with the variance of the BL probability for each normalization method in Figure 4-7 (upper).

This shows that the cases where the BL probability is most variable between replicates tend to be intermediate cases with BL probabilities closer to 0.5. And that if replicates show large variability this is usually independent of normalization method, which suggests the differences are caused by the sample or the experiment and couldn't be removed by normalization. It is also clear that Version_4 data (with improved mRNA treatment) generally gives a stronger BL signal (BL probabilities closer to 1.0), probably reflecting better experimental treatment of BL samples that can be significantly degraded. Finally it is clear that some of the larger variability between replicates occurs when one replicate is a tissue micro-dissection (performed with the aim being to enrich for tumour content and/or the most adequately

fixed area of the tissue), which would be expected to give stronger tumour specific expressions, and that this often leads to a clearer classification as BL.

In summary, the results from both versions show good reproducibility on the repeated samples, and the disagreement usually caused by the weaker signal of the arrays. Overall the classifier is able to effectively predict the samples according to their expressions, and better classification is achieved with improved array quality.

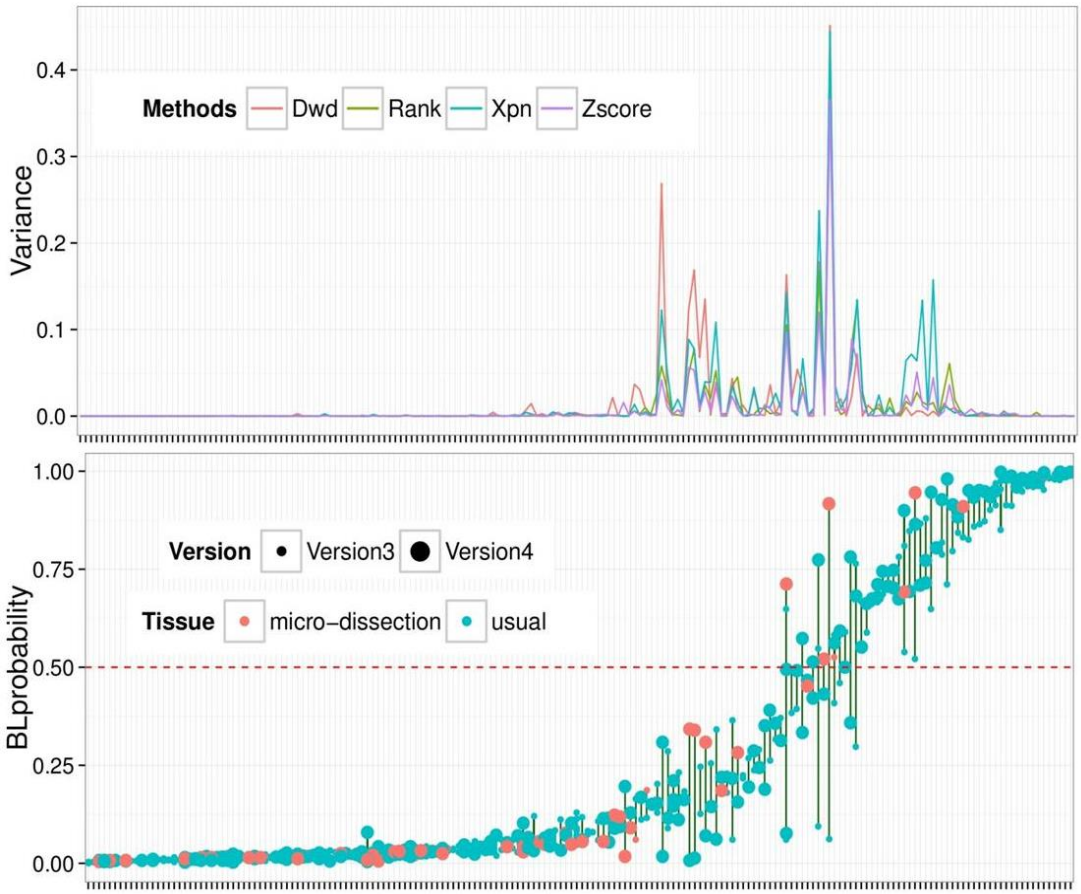


Figure 4-7: Classification reproducibility of the replicates from Version_3 and Version_4 data.

4.3 Concordance with diagnosis and clinical indication

The final gene expression classification was based on reducing the Illumina data set to a single replicate for each replicated case, and from above reproducibility analysis we came to the decision that choosing Version_4 data in preference to Version_3, micro-dissected tissue in preference to usual sampling, and otherwise choosing the newest array data. This gave a classification for 873 unique cases trained by GSE4475_0.95th data: 109 BLs and 764 DLBCLs, and we evaluated the concordance with the diagnosis made by the clinicians in HMDS.

4.3.1. Conventional diagnosis criteria

The current clinical diagnosis of these samples is based on a range of immunophenotypic and molecular FISH data, and FISH detected translocation are called rearrangement, so we also use this term when referring to the samples from the clinical patients. The detailed diagnostic pathway applied is described below.

First, if a case has a phenotype (the protein expression of the tumour cells carried out by flow cytometry and/or immunohistochemistry) of classic Burkitt's lymphoma, which usually displays CD20+, CD10+, BCL6+, BCL2-, Ki67=100% and deregulated P53 expression (P53+P21-), then FISH detection for *MYC*, *BCL2* and *BCL6* rearrangement is performed: 1) If it has *MYC* rearrangement and in absence of *BCL2* or *BCL6* rearrangements then this is diagnosed as Burkitt's lymphoma; 2) If it has *MYC* rearrangement and *BCL2* and/or *BCL6* then this is diagnosed as DLBCL with *MYC* rearrangement with additional comment of dual/triple hit; 3) If this has no *MYC* rearrangement this is diagnosed as DLBCL with additional comment of DLBCL with Burkitt phenotype and *MYC* negative.

Otherwise, if a case does not have a Burkitt phenotype then only FISH for *MYC* rearrangement detection is performed: 1) If it has *MYC* rearrangement, then it is diagnosed as DLBCL with *MYC* rearrangement, and FISH for *BCL2* and *BCL6* rearrangements detection are requested, where it has *BCL2* and/or *BCL6* at this point, it is commented with additional information as dual/triple hit, otherwise no *BCL2* or *BCL6* is commented as single hit; 2) If

MYC is normal it is just diagnosed as DLBCL with comments DLBCL *MYC* negative.

However as the samples are collected from different sources or clinical trials, there are cases that the diagnosis is not made locally by HMDS, and where the phenotype strongly suggest DLBCL the FISH detection for *MYC* is not performed. So in total, there are 72 cases diagnosed Burkitt's lymphoma, 48 DLBCL with *MYC* rearrangement, 753 DLBCL (285 are *MYC* normal 498 with *MYC* status unknown).

4.3.2. Concordance between GEP classifier and diagnosis

The agreement of the clinical diagnosis with the gene expression based classification is shown in Figure 4-8. Generally there is a high level of agreement between the two diagnoses (85% of cases diagnosed clinically as BL, and 95% of DLBCL), and about 30% of the DLBCL with *MYC* -rearrangement are classified as BL.

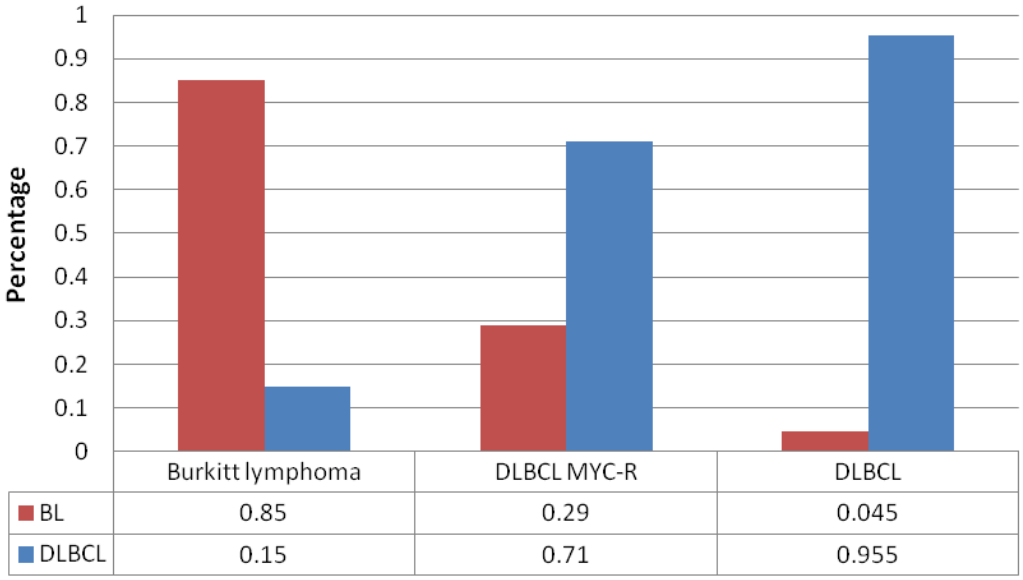


Figure 4-8: GEP classification comparison with current clinical diagnosis.

There are 11 clinical diagnosed BL cases that were classified as DLBCL and 34 DLBCL as BL by the GEP classifier. Of the 11 discrepant BL cases, 3 had classic BL characteristics, but the remainder of the group included a high level of aberrant cases: 3 have non-classic *MYC* arrangement, 5 exhibit a discrepant immunophenotype with the classic BL; and also there is 1 relapse case that might have developed different molecular features. The detailed characters of the clinical diagnosed BL while classified as DLBCL cases by the GEP classifiers are listed in Table 4-6. Figure 4-9 gives the occurrence of classic and non-classic *MYC* rearrangement in different BL probability intervals, and it shows that the proportion of non-classic *MYC* rearrangement is much higher in the low BL probability cases than that in high BL probability cases.

Table 4-6 : Characters of the diagnosed BL but GEP classified DLBCL cases

HMDS.ID	**BL-probability**	***MYC*-Rearranged**	**Phenotype**
1092	0.115	non-classic	BL Phenotype
H4151/08	0.357	non-classic	BL Phenotype
H2863/11	0.039	non-classic	BCL2+, P21+
1102	0.286	classic	expression of BCL2
H11436/06	0.206	classic	weak CD10 expression
H11920/10	0.313	classic	expression of BCL2
H1855/05	0.056	classic	relapse sample
H22569/11	0.455	classic	FOXP1+
H6237/13	0.407	classic	BL Phenotype
H811/06	0.343	classic	BL Phenotype
H8894/08	0.491	classic	BL Phenotype

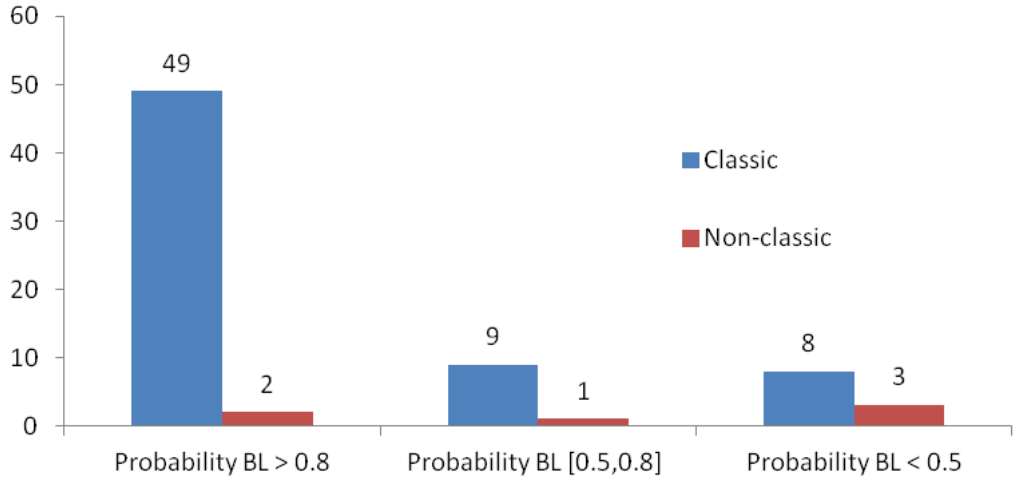


Figure 4-9: Number of classic *MYC* rearrangement in different BL probability intervals.

Of all 34 diagnosed DLBCL that classified as BL by the GEP classifier, 24 are from an ongoing clinical trial, for which other types of diagnosis features are not open yet. The clinical and survival status of the remaining 10 cases is listed in Table 4-7. Although the numbers are small, the table still shows some interesting findings of these GEP classified BL cases, 3 of the 10 cases show a BL phenotype even without MYC rearrangement, and the bottom 4 cases that have a relative low BL probability around 0.5 are all cured by R-CHOP treatment regardless of BCL2 or BCL6 rearrangement. However a similar case with a high BL probability 0.901 died when treated with R-CHOP, which suggests that R-CHOP maybe not an optimal treatment when the BL confidence is strong. It is hard to say if CODOX-M/IVAC is effective on high BL confidence cases by the limitation of number of cases.

Table 4-7: diagnosed DLBCL without MYC rearrangement classified as BL

HMDS.ID	BL-probability	Treatment	Survival-status	BCL2, BCL6 rearrangement
1066	0.805	CODOX-M/IVAC	Alive	
H3989/12	0.513	CODOX-M/IVAC	Alive	
H5288/11	0.747	CODOX-M/IVAC	Alive	BL phenotype
1039	0.951	CODOX-M/IVAC	Died	BL phenotype
1128	0.675	CODOX-M/IVAC	Died	BCL2 rearranged
H119/05	0.901	R-CHOP	Died	BCL2 rearranged
H5757/06	0.624	R-CHOP	Alive	BL phenotype
20317	0.538	R-CHOP	Alive	BCL2 rearranged
H15775/10	0.545	R-CHOP	Alive	BCL2, BCL6 rearranged
H19100/10	0.532	R-CHOP	Alive	BCL2 rearranged

4.3.3. Classification of the *MYC*-rearranged DLBCL

We also looked further at the small group that diagnosed as DLBCL but with *MYC* rearrangement detected. This is a group of particular interest, many of which are now referred as “lymphoma with features intermediate between BL and DLBCL”, though many studies have reported a poor prognosis currently there is no specific treatment for this group (more discussed in chapter 1 and chapter 5).

Of the 48 cases which were diagnosed in the DLBCL with *MYC* rearrangement category, about one third (14) were classified by GEP as BL and two thirds (34) were classified as DLBCL. However none of 14 classified BL cases has a BL probability greater than 0.8. This is plausible, as BL is a comparatively homogenous group with *MYC* translocation driven expression patterns, whereas in this group 39 of 48 cases also bear a BCL2 or/and BCL6 rearrangement in addition to MYC translocation, especially all 14

classified BL cases are concurrently BCL2 or/and BCL6 rearranged, which makes it a particularly complex group. The clinical and survival data of the cases classified as BL and DLBCL are listed in Appendix B tables.

We analysed the 35 cases treated with R-CHOP and have survival status available, of which 10 cases were classified as BL and 25 as DLBCL by the GEP classifier. And surprisingly, in the BL category 70% (7 out of 10) died or had a persistent response to the treatment, while in the DLBCL category the 32% (8 out of 25) experienced death/persistent response, which have no significant difference with the *MYC*-negative DLBCLs (p=0.4, comparing 19 *MYC*-rearranged and 106 *MYC*-normal samples treated with R-CHOP). Although these numbers are small, it implies that the intermediate cases classified as DLBCL by gene expression have no significant difference in response to R-CHOP treatment than other DLBCLs, while the cases classified as BL by gene expression have worse prognosis compared to normal DLBCL diagnosed by conventional criteria.

In addition, the analysis between the single-hit (*MYC*-rearranged in the meantime both *BCL2* and *BCL6* normal) and double-hit (*MYC*-rearrangement with *BCL2/BCL6* rearrangement) cases reveals that single-hit cases do not have superior survival compared with double-hit cases (35 RCHOP treated, survival rate of single-hit and double-hit are 37.5% and 63% respectively), at least from this dataset.

In conclusion, the result shows that the classifier can be successfully applied to our own clinical FFPE samples from routine practice with good concordance with more traditional diagnostic methods, and produced data that suggests possible prognostic value (on standard RCHOP treatment) for those intermediate cases classified as BL by the gene expression classifier but DLBCL by our more standard methods.

4.4 R package implementation

The GEP classifier investigated in the above studies is implemented in an R package BDC (Burkitt lymphoma and Diffuse large B-cell lymphoma classifier), using the R package mechanism [190], and is available for public

use from the GitHub website https://github.com/Sharlene/BDC. The package provides a list of options discussed in previous work, including classifier gene set, cross-platform normalization method and training dataset thresholds, with the default setting as 28-gene, Z-score normalization and trained by GSE4475 0.95 threshold.

Users can download the BDC package and install it on their own R environment. This is an easy to use classifier with only one main function, *classify*. Users need to provide the gene expression of the test samples after self normalization within their own platform, and the gene symbols for each probe, however it doesn't have to be unique or update to the latest gene name, the classifier would merge the replicate genes and check the gene names automatically. Also the package provides a simple summary and plot functions for users to interoperate the classification results. An example of using the classifier is illustrated in Figure 4-10.

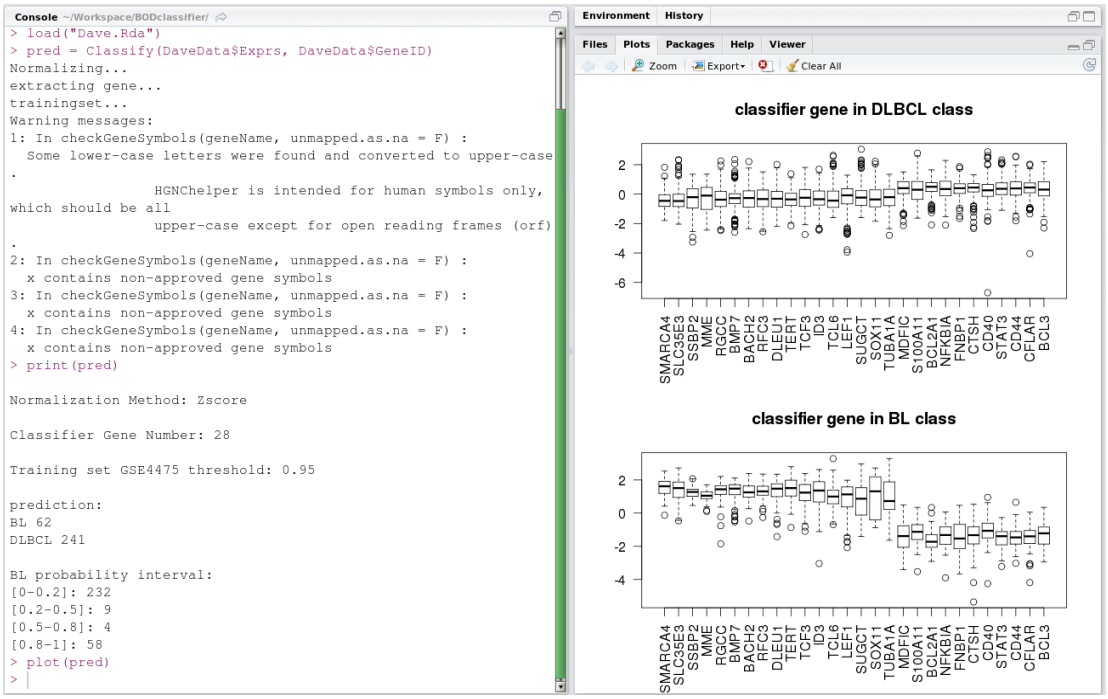


Figure 4-10: An example of the classifier implemented in an R package BDC.

4.5 Conclusion

In this chapter, we examined the capability of the GEP classifier developed in the chapter 3 on the FFPE samples collected by collaborators. The comparison of different normalization methods and training set options indicate that DWD method may not be effective on FFPE DASL arrays, and that the classifier trained by GSE4475 threshold of 0.95can better recognize those BLs with weaker signals. In addition, the repeat samples on Version_3 and/or Version_4 show high reproducibility within and between platforms, and the cases express large variance are usually explained by the quality of the specific sample.

Moreover, the classification results of the 28-gene LibSVM classifier with Z-score normalization and GSE4475 threshold of 0.95 on the FFPE samples are highly concordant with the diagnosis made by clinicians using conventional methods: 85% agreement on BL cases and 95% on DLBCL. The classification result indicates that most DLBCLs and BLs have different expression pattern but there are cases lying in-between with lower confidence of which class it belongs to. The classifier classifies about 30% of the clinical diagnosed DLBCL with *MYC* rearrangement cases as BL, and the survival status of the 35 RCHOP treated cases suggest a worse repose of the classified BL cases comparing to classified DLBCL cases. The work presented in this chapter and last chapter have been accepted for publication in Genome Medicine.

Chapter 5
The role of *MYC* in non-Burkitt lymphoma

The objective of the work carried out in this chapter is to gather as many resources as possible to investigate important topics related to *MYC*-associated non-Burkitt lymphoma (mainly DLBCL and BCLU in our study): (1) whether we can identify a gene expression pattern that predicts *MYC* translocation, (2) to what extent *MYC* abnormality correlates with patients outcome in non-Burkitt lymphoma, and (3) to discover *MYC*-associated genes and potentially involved molecular functions/pathways corresponding to different subtypes. The results are presented in sections 5.1 to 5.3 respectively.

5.1 *MYC* translocation expression pattern

MYC translocations are usually identified in clinical samples using fluorescence *in situ* hybridisation (FISH). There is also a group who successfully identifying *MYC* translocations by testing nuclear MYC protein with an immunostaining and screening method [191]. Here we investigate if the translocation is associated with a characteristic pattern of gene expression in non-BLs, and whether this pattern can accurately identify samples with *MYC* translocations, thus eliminating the need for these assays.

This is done by fitting a classifier on the public available dataset that has FISH detected *MYC* status, and testing the validity on our in-house HMDS datasets. Considering the potential distinct mechanism behind *MYC* translocation in different lymphoma types, we also explore this topic with the consideration of each DLBCL subtype: ABC, GCB and UCL respectively. This is examined by comparing gene expression between the cases with and without *MYC* translocations, and selecting a set of significantly differentially expressed genes between the two *MYC* groups. Then we explore if the expression pattern of these differential genes can identify the *MYC* translocation by generating a *MYC*-translocation classifier. The

classifier is built and optimized on datasets downloaded from the NCBI database and the prediction performance is evaluated on HMDS DASL array datasets generated by collaborators (DASL array data used in chapter 4).

5.1.1. Datasets summary

At the stage of developing *MYC* translocation classifier, we gathered the gene expression profiles of only fresh-frozen non-BL samples that have detected *MYC* status from the Gene Expression Omnibus (GEO), because *MYC* translocation in non-BL usually occurs in a complex cytogenetic context, and paraffin embedded data may add extra noise that make it more difficult to compare between groups. There are 222 samples included in this step: a subset from GSE4475 which contains 172 non-mBL and intermediate cases; cases from GSE44164 consisting of 32 *MYC*-translocated DLBCL; and an additional 18 *MYC*-translocated cases from a 271 DLBCLs dataset GSE22470. The detail is given in Table 5-1. All together there are 85 *MYC* translocated cases and 137 *MYC* negative cases. Since all datasets were performed on Affymetrix U133a GeneChip, we combined three raw datasets as one and referred it as the MYC dataset in the following study, and then we processed the dataset with the affy package for background correction, normalization and probe intensity extraction as done in chapter 3.

Table 5-1: Datasets used in developing *MYC* translocation signatures

Data set	Group	*MYC*-translocated	*MYC*-negative
GSE4475	Hummel[129]	22 GCB, 10 ABC, 3 UCL	56 GCB, 46 ABC, 35 UCL
GSE44164	Valera A[45]	28 GCB, 1 ABC, 3 UCL	
GSE22470	Salaverria[192]	10 GCB, 3 ABC, 5 UCL	
MYC data	Combined	60 GCB, 14 ABC, 11 UCL	56 GCB, 46 ABC, 35 UCL

Then several SVM classifiers based on selected *MYC* translocation signature gene sets were compared and tuned on the MYC dataset by a cross-validation method. The optimized classifier was next validated on our FFPE DASL data. There are 476 samples with FISH detected *MYC* status in version 3 and version 4 data sets together, and we excluded the samples with BL probability over 0.8, which are very confident BLs (because we're

exploring the *MYC* expression pattern in non-BL cases). This left us 69 rearranged and 280 negative cases. In addition, we also predicted the *MYC*-translocation status of other three large public DLBCL datasets with the classifier. However there is no FISH detected *MYC* status available for those datasets, and we compared the survival difference between the predicted *MYC*-rearranged and *MYC*-negative groups. Details of the validation datasets are listed in Table 5-2.

Table 5-2: Datasets used to validate MYC translocation signatures

Data set	Sample description[1]	Tissue[2]	Treatment and Survival[3]
GSE4732(1)[128]	54 BL and 249 DLBCL, *MYC* unknown	FF	157 CHOP
GSE10846[193]	420 DLBCL, *MYC* unknown	FF	181 CHOP; 233 R-CHOP
GSE31312[194]	498 DLBCL, *MYC* unknown	FFPE	470 R-CHOP
DASL data	69 rearranged; 280 negative	FFPE	152 R-CHOP; 71 CODOX-M/I-VAC

[1]MYC unknown means no detected MYC status available in original dataset.

[2]FF denotes fresh frozen, and FFPE denotes formalin fixed paraffin embedded tissue

[3]List the number of cases both have treatment and survival follow up information available.

5.1.2. Differentially expressed genes

We first tested whether there is an obvious distinct expression pattern in *MYC*-rearranged cases compared to *MYC*-negative cases without considering subtypes. This is done by fitting a simple two group comparison linear model with the limma package. Next supported by the hypothesis that *MYC* activities are different in different cancer types, and the fact that *MYC* translocation were predominately found in GCB subtype, we then added in the factor of three subtypes ABC, GCB, UCL, to see how *MYC* translocation differs inside each subtype. It could also avoid the potential drawback that genes are actually selected due to the differences between subtypes. This is also carried out with limma package, by fitting a more sophisticated two factors interaction linear model.

In each comparison model, differently expressed probes are selected by setting the threshold at the adjusted p value smaller than 0.001 and the absolute log fold change greater than 0.5. An example of some other detailed statistic values for the probe selection in the simple two group comparison is illustrated in Figure 5-2.

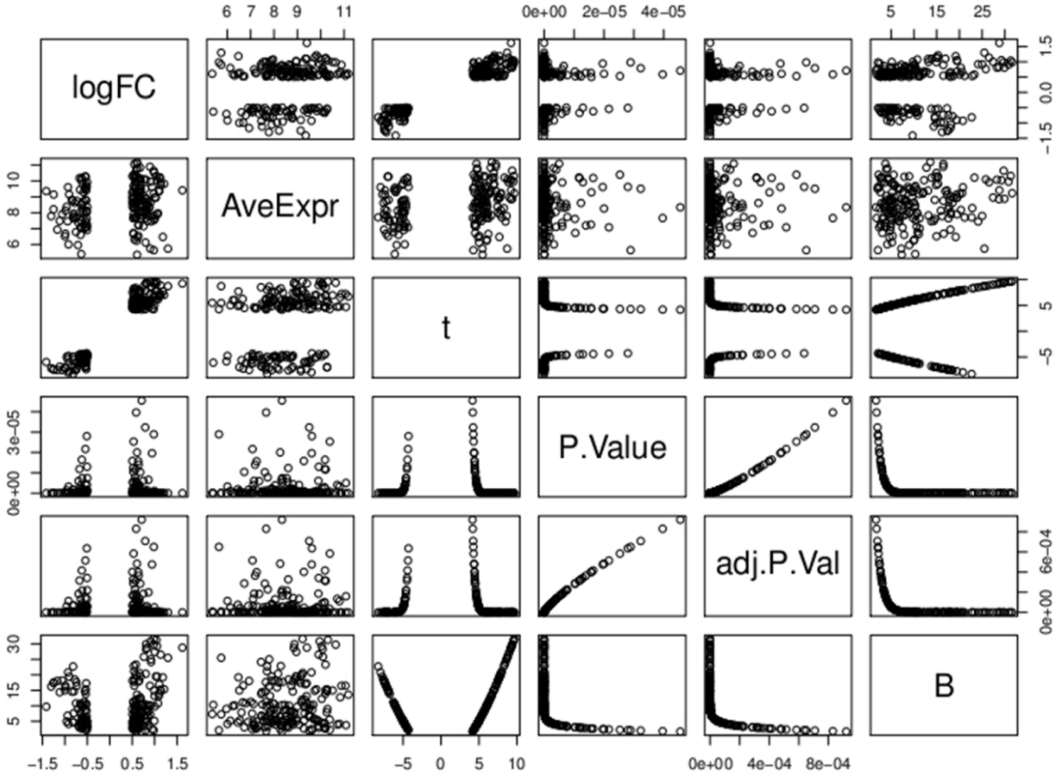


Figure 5-1: **Differentially expressed probed between *MYC*-rearranged and *MYC*-negative non-BLs.**

There are 192 probes picked out as significantly differentially expressed (67 up-regulated and 125 down-regulated) in the simple two groups comparison (all subtypes together). However there are fewer probes fulfil the selection criteria when considering subtype effects, especially in ABC and UCL (see the detail in Table 5-3). This may be because there are less *MYC*-rearranged cases in those subtypes (14 and 11 respectively), or some other reasons that obscure the expression of *MYC* translocation. Even fewer probes come out as significantly differentially expressed when we compare

the two *MYC* groups with only ABC or UCL samples (not shown in the data). Figure 5-3 gives the relationship of the probes selected under each condition. It shows that probes selected within each subtype have a rather small overlap, and even a considerable number of probes picked by all types were not selected as differentially expressed in any subtype. There are a small number of 15 probes consisting of 1 commonly up-regulated gene *MYC* and 11 down-regulated *JAK1, PTPN1, AHR, BCL3, CDK5R1, RASGRP1, BATF, STAT3, CFLAR, SOCS1* overlapping by all three subtypes.

Table 5.3: Number of selected genes under each condition

	All type	ABC	GCB	UCL
Down-regulated probes	125	21	83	44
Up-regulated probes	67	25	56	28
Down-regulated genes	84	18	60	33
Up-regulated genes	50	22	40	20

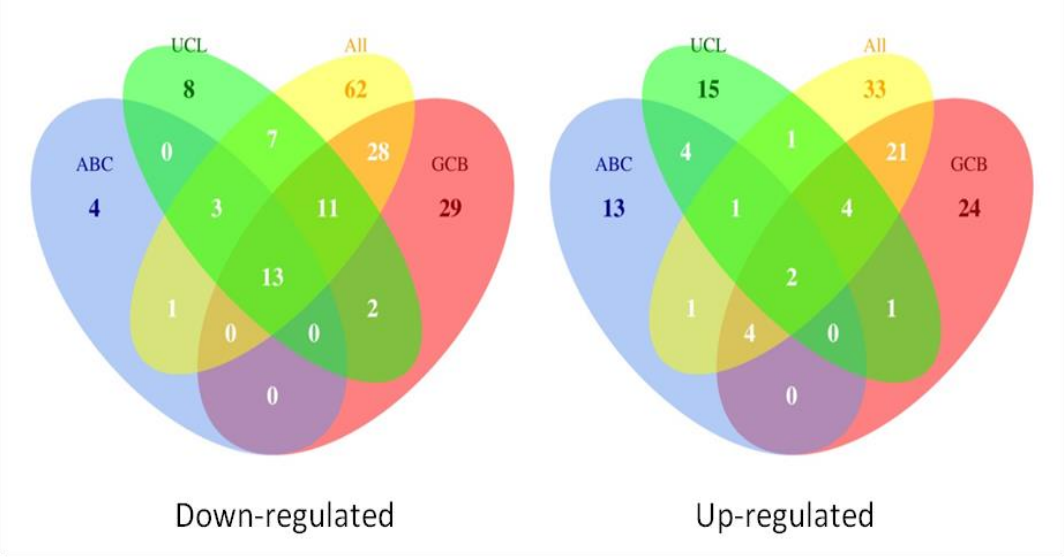


Figure 5-2: Venn diagram of differentially expressed probes between *MYC* translocated and *MYC* normal samples in each situation.

5.1.3. Predict MYC translocation by gene expression

To examine how the signature genes selected above can be predictive of *MYC* translocation, we first tested if they can successfully recognize *MYC*-rearranged samples in the *MYC* data by constructing a SVM classifier for each gene set. The classifiers were evaluated by a 10-fold cross-validation method, and the predicted result is presented in Table 5-4. It shows that generally the *MYC* status can be accurately predicted with the gene set selected by all types with the sensitivity around 80~90% in rearranged cases and >95% in *MYC*-negative cases. It is no surprise that the best prediction within each subtype is given by the probes selected from the corresponding subtype, while the accuracy drops dramatically when predicting with probes selected within other subtypes (50~70%). And the probes selected from all types can predict the *MYC* translocation status of each subtype most close to the best result that achieved by the probes selected from this subtype.

Table 5.4: Classification results of classifiers built with each gene set

	FISH status	**Gene.All**	**Gene.ABC**	**Gene.GCB**	**Gene.UCL**
MYC-R[1] cases prediction in each subtype					
ABC subtype	14	11(79%)[2]	12(86%)	8(57%)	7(50%)
GCB subtype	60	55(92%)	47(78%)	57(95%)	34(57%)
UCL subtype	11	9(82%)	8(73%)	7(64%)	9(82%)
MYC-N[1] cases prediction in each subtype					
ABC subtype	46	46(100%)	46(100%)	43(93%)	45(98%)
GCB subtype	56	53(95%)	43(77%)	54(96%)	53(95%)
UCL subtype	35	35(100%)	32(91%)	33(94%)	35(100%)

[1] *MYC*-R denotes *MYC*-rearranged and *MYC*-N denotes *MYC*-negative cases.

[2] The percentages are the accurately recognized cases (sensitivity).

We then assessed the four lists of genes on DASL version 3 and version 4 datasets, an additional randomly selected 100 genes was also assessed as control set. Similarly, gene sets were used to build SVM classifiers, however the classifiers were trained on the *MYC* dataset and tested on DASL data.

Figure 5-3 shows the correlation of the predicted BL probability of each SVM classifier for the DASL Version_3 and Version_4 data sets. We can see that the correlation of the BL probabilities is high (0.84 ~0.92) among the classifiers built with the four gene lists selected above and low (0.55~0.65) with the random selected genes, which suggests the selected genes predict the result based on biological meaning.

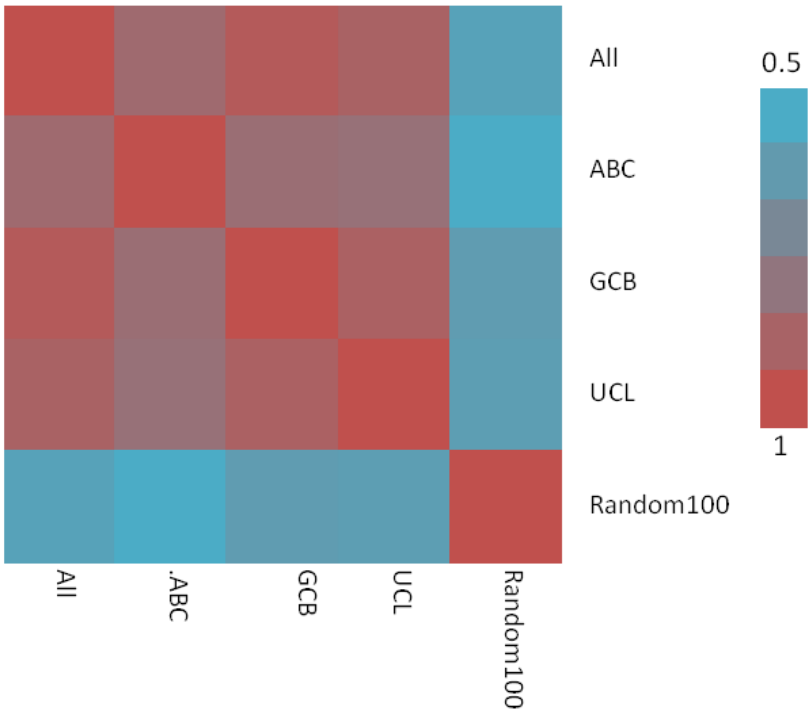


Figure 5-3: Correlation of clinical FFPE samples *MYC* prediction confidence classified by different lists of genes.

We took the result of the classifier built with the 134 genes (192 probes) selected in all types as the final predicted *MYC* rearrangement status, and merged the replicate samples in DASL data as done in chapter 4 (Section 4.2.3). The comparison of FISH detected *MYC* translocation with *MYC* status predicted by gene expression classifier is showed in Table 5-5. The result shows that after excluding the confident BL (BL classifier probability > 0.8), 19 of the 280 samples that don't have FISH detected translocation are

predicted as *myc-r* (*MYC* rearranged), while over 40% (31 from 69) of the detected rearranged cases are not correctly recognised. We also tested the classifier on the confident BL samples (which all bear *MYC* rearrangement), and 50 of 51 cases are correctly predicted as *myc-r* (see Figure 5-5).

Table 5-5: Confusion matrix of the *MYC* rearrangement classifier on DASL data

	Predict *myc-r*	**Predict *myc-n***	Sensitivity[1]
FISH detected *MYC-R*	38	31	0.55
FISH detected *MYC-N*	19	261	0.93
Specificity[1]	0.67	0.89	

[1]Sensitivity and Specificity for each category is calculated by taking the correctly classified cases as true positives, for example *MYC-N* category, the TP is predicted *myc-n* cases, and FP is predicted *myc-r* cases.

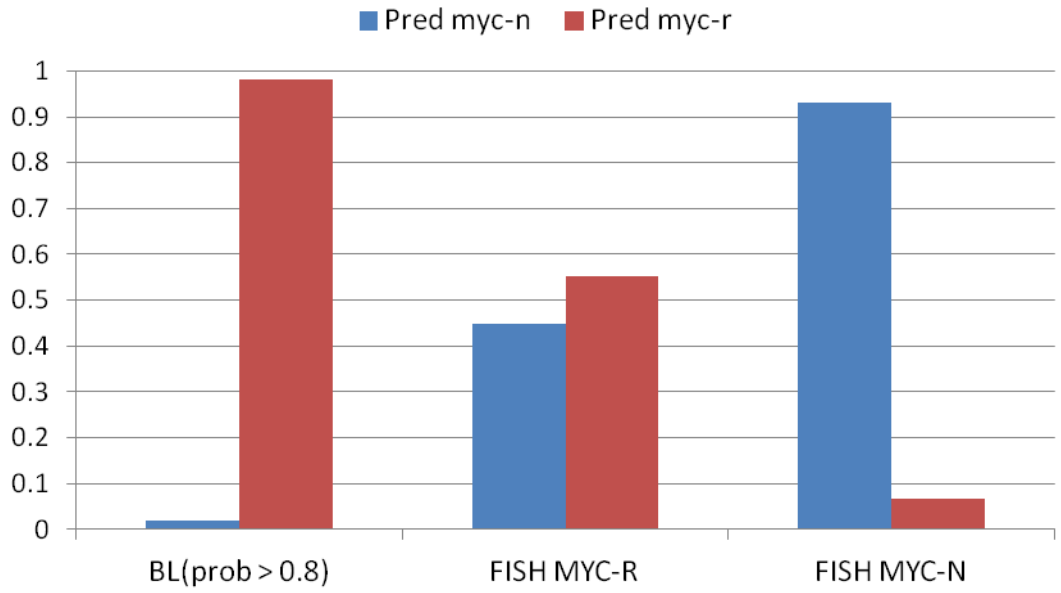


Figure 5-4: Comparison of original diagnosis and *MYC* FISH detection versus GEP BL classification and *MYC* status prediction

So the classifier can identify *MYC* rearrangement in the definite BL samples, but is less effective in BCLU and DLBCL cases (less confident BL). This was also shown in the public datasets that we used to develop the classifier, not all *MYC* rearranged samples were correctly classified (sensitivity is around 80~90%). And this is even more obvious in the DASL data. However it is interesting to see that although the samples are detected to bear *MYC* rearrangement by FISH, the *MYC* mRNA expression is significantly higher in the predicted *myc-r* class than that in the *myc-n* class. This is shown in Figure 5-6. The potential reasons for this may be: (1) there are cases of MYC rearrangement that don't show the associated expression pattern, or the expression is compromised by other aberrations (*BCL2, BCL6*), (2) the translocation expression pattern can caused by other reasons like *MYC* mutations, gains, and (3) it could also be a problem of transferring the classifier to the new platform and the noisier FFPE data.

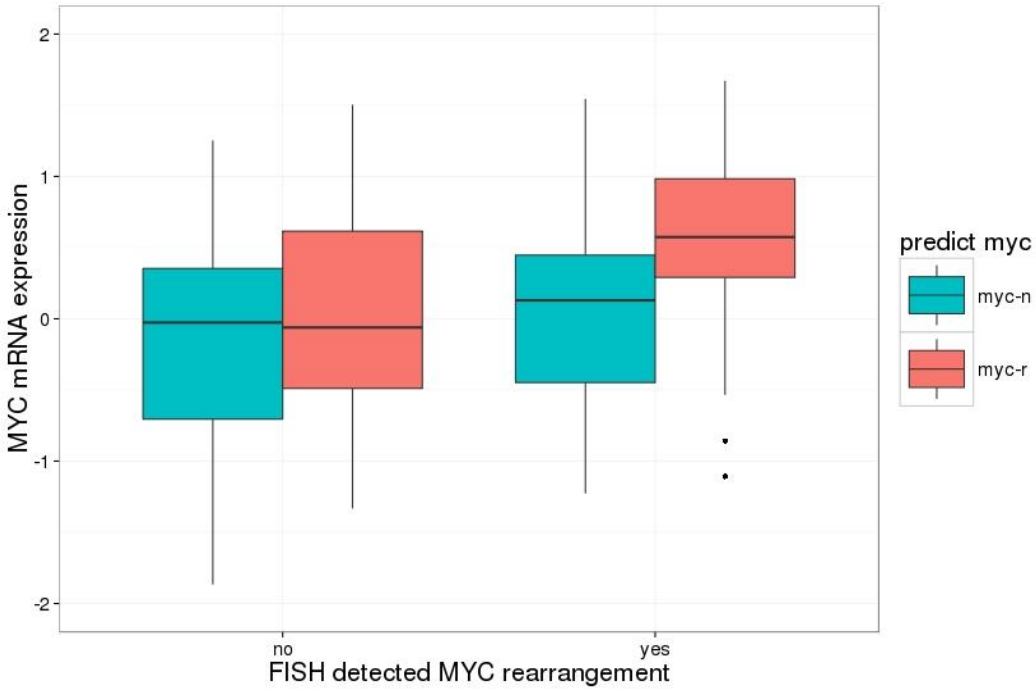


Figure 5-5: *MYC* mRNA expression in FISH detected and predicted *MYC* rearrangement groups of DASL data.

Although *MYC* rearrangement expression pattern doesn't seem to be in full concordance with FISH detected *MYC* status, it is possible that the expression pattern reflects the biological mechanism and is responsible for the inferior outcome of this *MYC*-associated lymphoma. So we next test the above classifier on other three GEO datasets: GSE4732_P1, GSE10846 and GSE31312. In GSE4732_P1, again all BLs were predicted as *myc-r* group. However the detected *MYC* status is not available in the DLBCL cases or the other two datasets, and we explored the prognostic impact of *MYC* rearrangement expression pattern with the treatment and survival information on the non-BL patients. This is showed in section 5.3.

5.2 *MYC* impact on survival

The impact of *MYC* activity on patient outcome is elusive, though reported in many studies: translocation and other type of *MYC* mutations, high level of *MYC* gene expression, protein over-expression together with a BCL2 effect all may contribute to poor prognosis. Here in this chapter, we first assessed the survival difference between two groups of patients: *MYC*-rearranged and *MYC*-negative according to the prediction of above *MYC* translocation classifier. Secondly, we evaluated the effect of MYC mRNA expression level on patients' survival. Additionally, we checked the *BCL2* mRNA expression level impact on survival by itself as well as in combination with *MYC* mRNA expression and other prognostic factors. The data used consists of the DASL data which has clinical information (treatment, survival and follow-up time) available, and also we performed a retrospective analysis on two large datasets GSE10846 and GSE31312 as well as the DLBCL samples in GSE4732_P1.

5.2.1. Survival impact of GEP predicted *MYC* translocation

In DASL dataset, there are 152 R-CHOP treated and 71 CODOX-M/IVAC treated samples. However the Kaplan-Meier survival estimator didn't show significant difference between the predicted *myc-r* and *myc-n* patients, neither in R-CHOP treated cases ($p = 0.383$) nor CODOX-M/IVAC treated cases ($p = 0.737$). Moreover there is no significant difference between FISH detected *MYC* rearrangement and *MYC* negative cases either, with RCHOP

treated p value 0.102 and CODOX-M/IVAC treated p value 0.307. The fact that neither FISH detected *MYC* rearrangement nor predicted *myc-r* expression pattern show a significant worse outcome may suggest there are some other factors (e.g. age, disease stage) adding the complexity of the prognosis.

A similar situation appears in the other GEO datasets: no significant survival difference is observed between predicted *myc-r* and *myc-n* cases. However it shows a significant separation in RCHOP treated GCB subtypes (see Table 5-6 and Figure 5-7). This may due to the fact that *MYC* translocation predominately occurs in GCB subtype, and that the *MYC* translocation expression pattern pulls out the cases respond sub-optimally to the treatment.

Table 5-6: Survival difference between predicted *myc-r* and *myc-n* groups: p-values assessed by Kaplan-Meier model in different treatments and subtypes

	All	**ABC**	**GCB**	**UCL**
GSE4732-CHOP	0.828	0[1]	0.411	0[1]
GSE10846-CHOP	0.683	0.117	0.314	0.011
GSE10846-RCHOP	0.638	0.307	0.022	0.628
GSE31312-RCHOP	0.375	0.356	0.003	0.82

[1]Survival difference equals to 0 because there is no predicted *myc-r* case

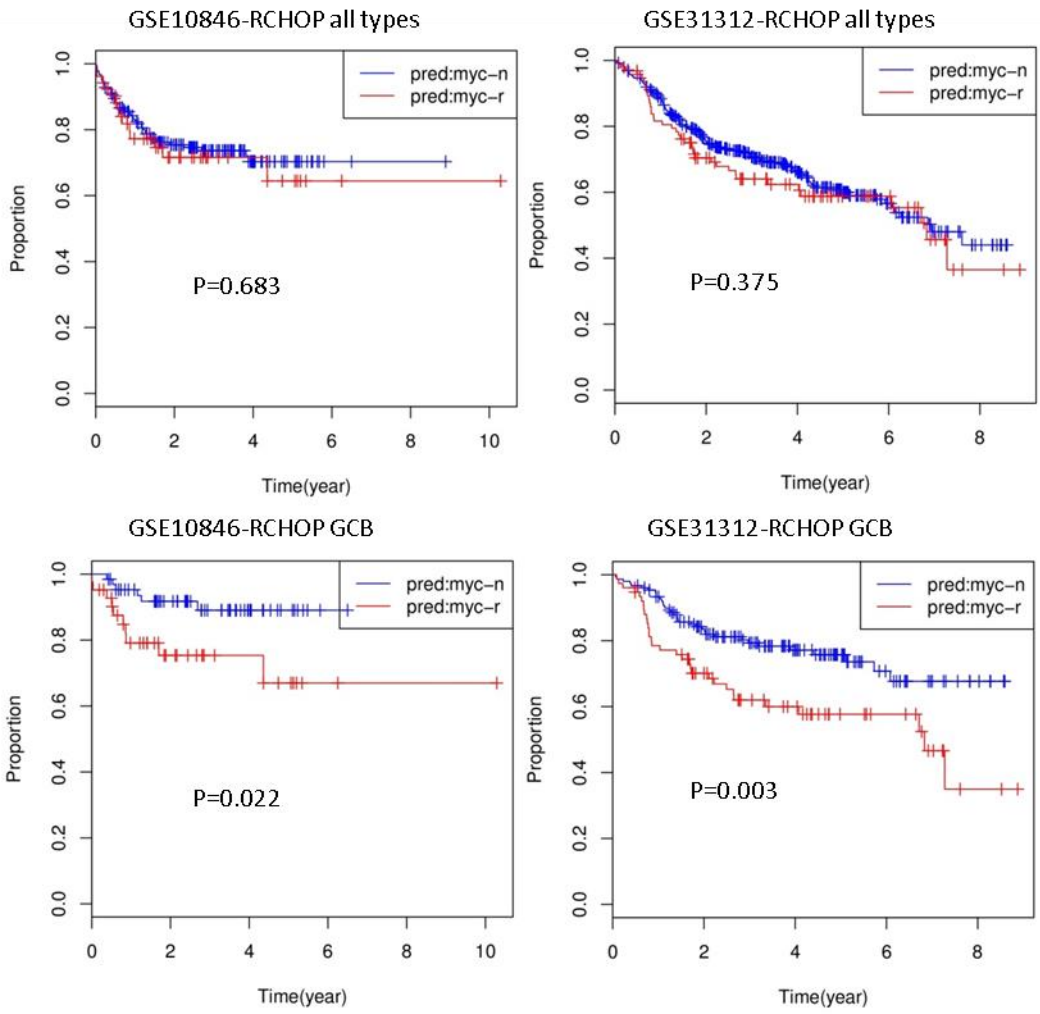


Figure 5-6: Kaplan-Meier survival curve in GES 10186 and GSE31312 RCHOP treated cases

5.2.2. Survival impact of *MYC* mRNA expression

Next we checked the *MYC* expression effect on patients' survival in three above published datasets. As *MYC* expression is a continuous value, we first checked the correlation of the expression and the estimated survival time by Kaplan-Meier model. This is calculated by counting the number of pair-wise agreements/disagreements between two ranking lists. In the context, agreement is defined as two random selected objects where the observation with the shorter survival time of the two also has the higher *MYC*

expression, and disagreement is the other way around. It computes all $n(n-1)/2$ pairs of data points, and the concordance equals to $(agree + tied/2)/(agree + disagree + tied)$. The concordance of *MYC* expression and survival time in the three public datasets is illustrated in Figure 5-8. The values are around 0.6 under each subtype and treatment, while GCB subtype especially with R-CHOP treatment has the largest concordance. This is also consistent with what we found between treatment response and the predicted *MYC* translocation cases.

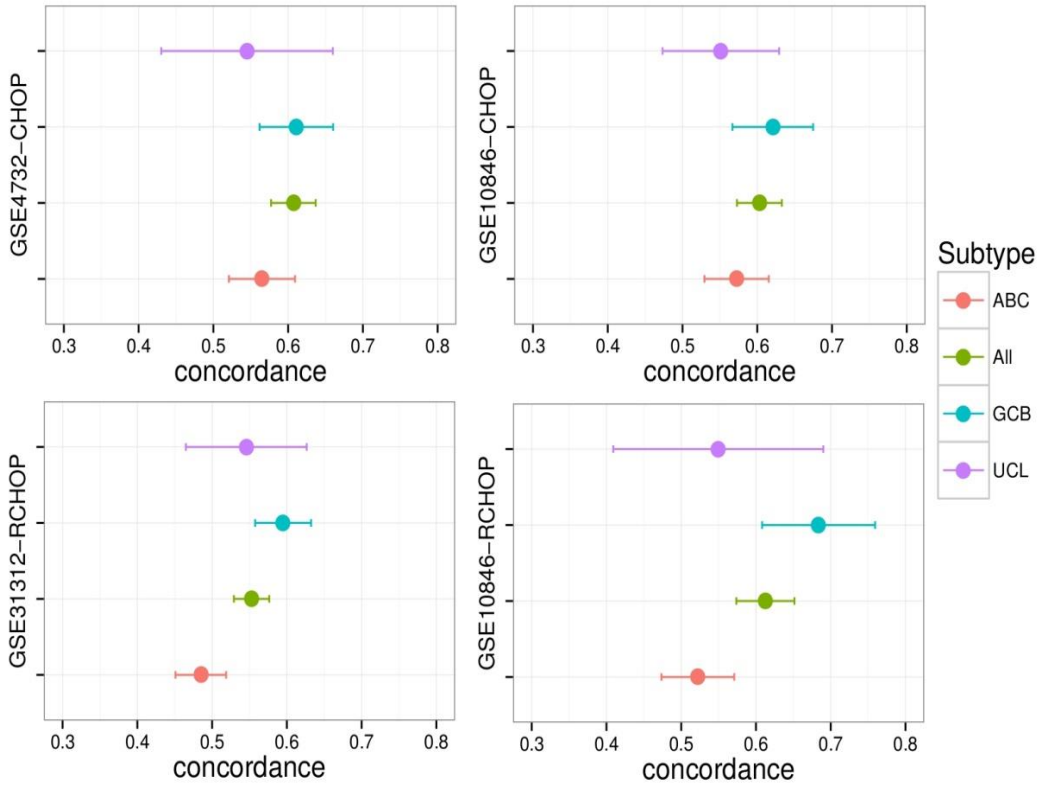


Figure 5-7: Concordance between *MYC* expression and follow up time in each subtype.

Next we divided the samples in each dataset into four categories: very high, high, low and very low by setting the boundaries at the 20th, 50th and 80th centiles of the *MYC* expression ranked list, and evaluated the survival differences by Kaplan-Meier model. Figure 5-9 shows the estimated survival curve for every category in each dataset. A significant separation is observed in all datasets. The very high expression of *MYC* is strongly correlated with poor survival. The CHOP treated data set GSE4732 shows the clearest survival separation on the basis of *MYC* expression. The RCHOP treated GSE31312 shows the expected improved overall survival attributed to the better treatment, and also an obvious survival separation between the very high *MYC* expression and the rest groups, however this is less pronounced among other groups. Dataset GSE10846 which contains both CHOP and RCHOP treated cases shows similar survival separation by *MYC* expression in both treatment groups.

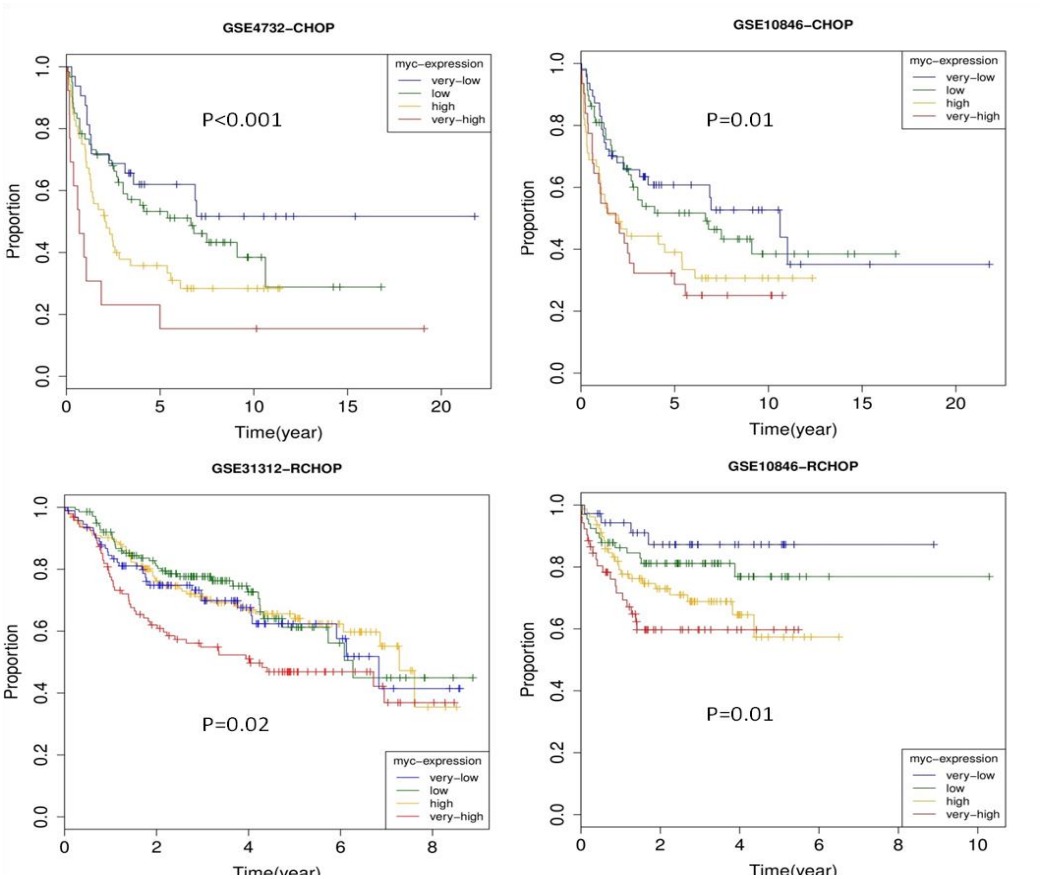


Figure 5-8: Kaplan-Meier survival curves of four *MYC* expression categories in each dataset.

5.2.3. Survival impact of *MYC* mRNA combined with other factors

As discussed above, the prognostic impact of *MYC* expression in cooperation with other potential factors, especially its correlation with *BCL2* is not clear. Here we first investigated the *BCL2* mRNA expression as a single effect on the patients' survival. Similarly it was divided into four categories (very high, high, low, very low), and the survival curves of each category for each dataset is illustrated in Figure 5-10.

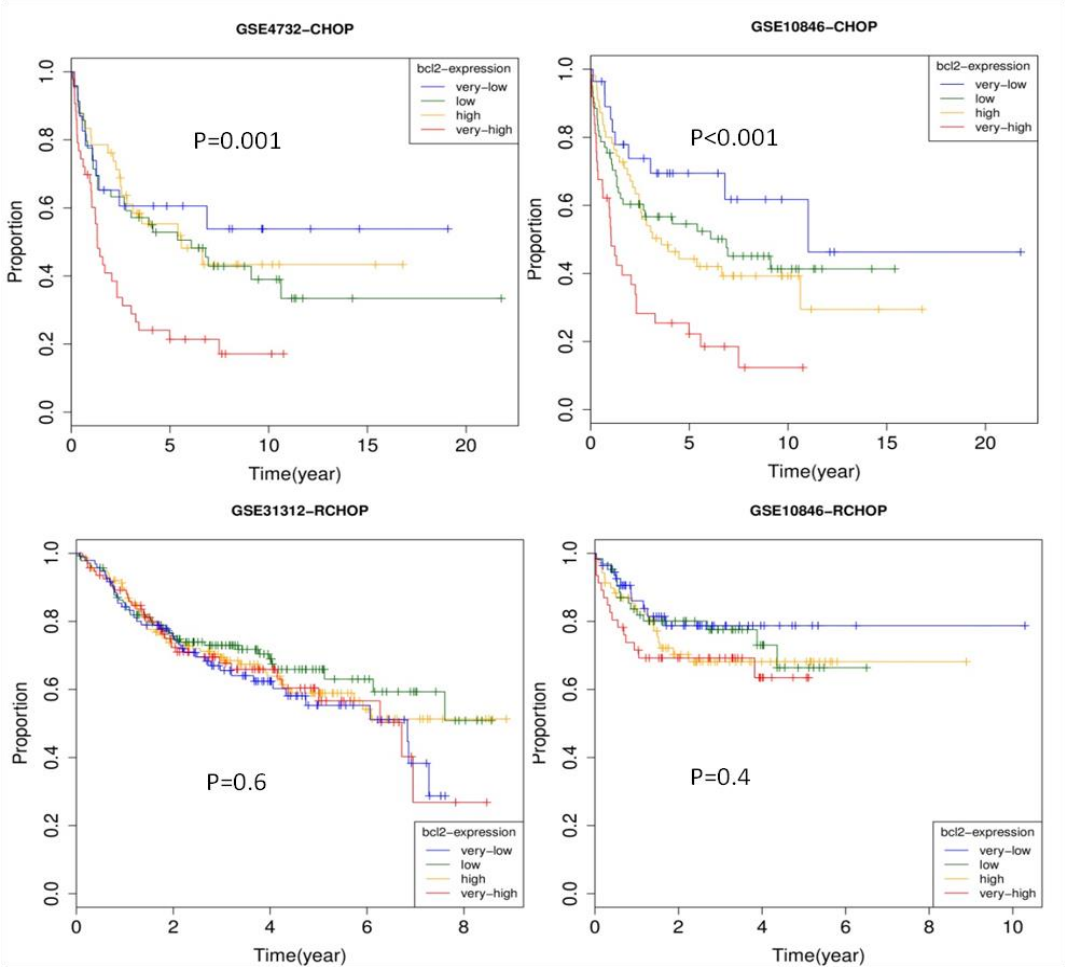


Figure 5-9: Kaplan-Meier survival curves of four *BCL2* expression categories in each dataset.

The figure shows that high *BCL2* expression is significantly correlated with short survival (p-value ~ 0.001) on CHOP treated cases, but no obvious effect (p-value > 0.4) was observed when patients are treated with RCHOP

regimen. Therefore it is possible that the significant difference on CHOP treated cases is caused by the overall poor response to the regimen, and when the treatment developed the BCL2 expression effect disappears. So *BCL2* mRNA expression itself is not associated with worse outcome in RCHOP treated patients.

To further investigate the co-impact of *MYC* and *BCL2*, we divided both *MYC* and *BCL2* expression into two groups: high and low, samples with expression values higher than 50% all samples are referred as high otherwise as low. The *MYC* and bcl2 co-effect is shown in Figure 5-11.

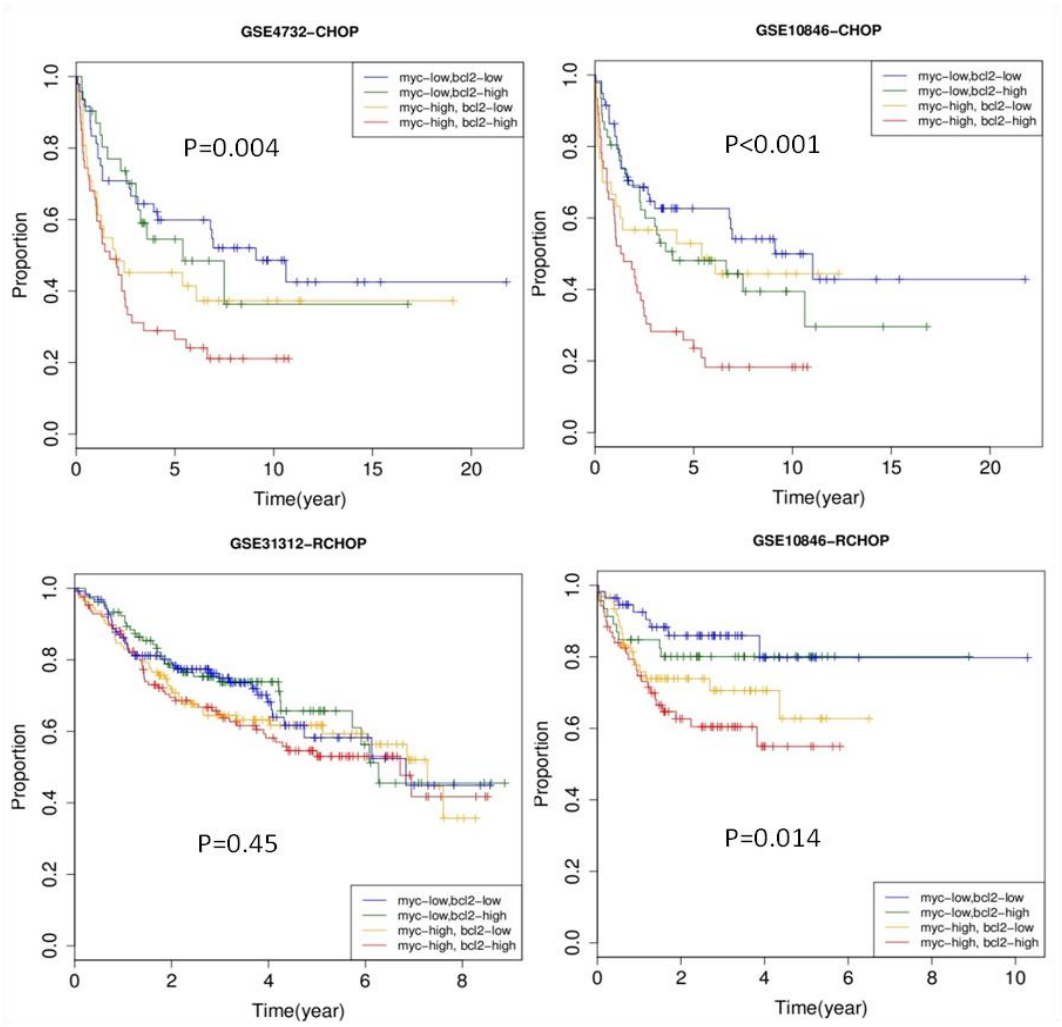


Figure 5-10: Kaplan-Meier survival curves of *MYC* and *BCL2* co-expression categories in each dataset.

In all datasets except GSE31312 exhibit a separation of different groups and that high expression of both *MYC* and bcl2 has the worst survival, while low expression of both has the relatively longest survival. As the dataset GSE31312 is generated from FFPE samples, it is possible that the simple half cut-off on the *MYC* and *BCL2* expression is not sufficient for the analysis. Also in the CHOP treated datasets, we can see *BCL2* and *MYC* expression level play similar impact on the survival, with no obvious different between the groups that has one expression high while the other low. However, the impact of *BCL2* drops vastly in RCHOP treated dataset GSE10846, with a clear strong inferior effect of *MYC* high expression, although high level of *BCL2* still suffer worse survival compared with that of low level.

A recent study suggested that the co-expression of MYC and BCL2 protein is the basis of poor outcome in all subtypes of DLBCL, and is a better survival indicator than the GEP subtypes [195]. When we check this from mRNA expression, it does seem that high expression of *MYC* and *BCL2* is predominantly seen in ABC subtype (see Figure 5-12), which is known to have worse outcome. To further explore this, we investigated the co-expression level of *MYC* and *BCL2* in ABC and GCB subtypes respectively, and the differences of survival influence among the four co-expression level categories on are listed in Table 5-7.

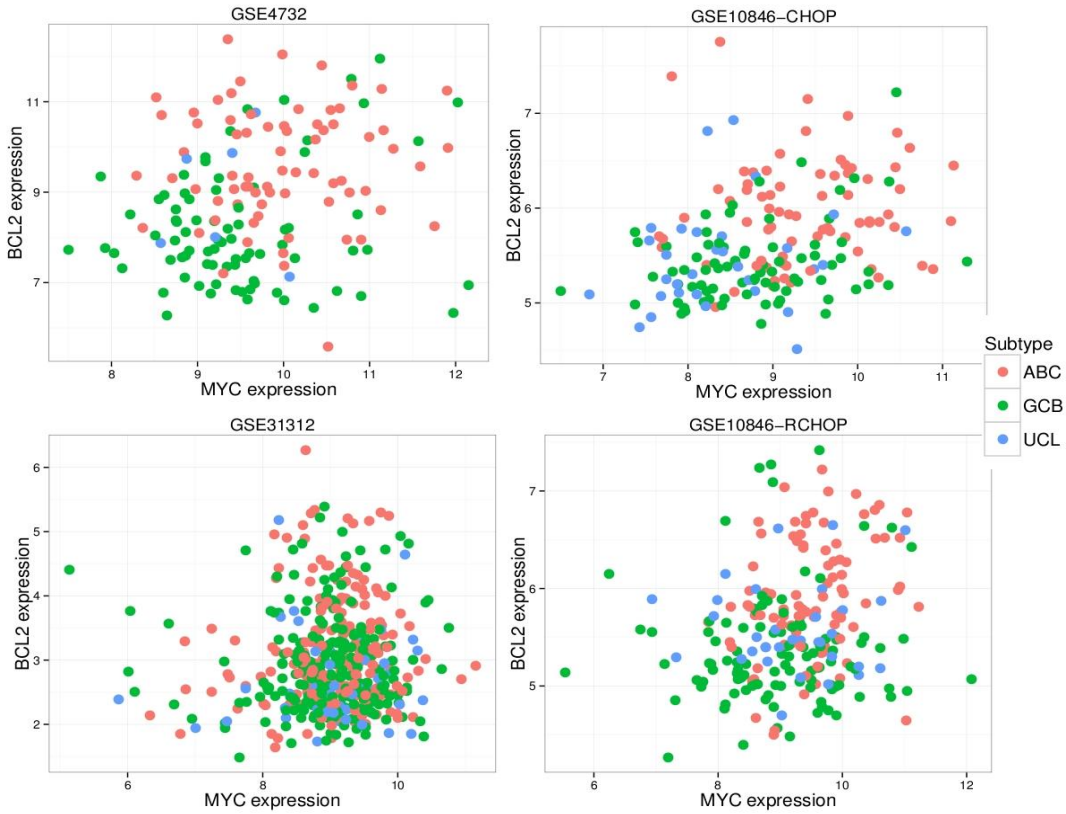


Figure 5-11: *MYC* and *BCL2* mRNA expression in different DLBCL subtypes.

The table shows that there is no difference among the four expression level groups in ABC cases in any dataset, and that there is significant difference in CHOP treated GCB cases but not RCHOP treated cases. The fact that the significant survival difference among different expression level exist when considering all subtypes while no significance different difference showing out within a particular subtype indicates the overall poor prognosis in not fully explained by high level of *MYC* and *BCL2* mRNA expression. We also investigate the co-impact of *MYC* and *BCL2* by examining the *MYC* expression effect (high and low group with a cut-off of 50%) within two *BCL2* expression categories (high and low with 50% cut-off). The result shows that there is still a significant difference in the RCHOP treated GSE10846

dataset, which is the most informative one (better sample quality and modern treatment). This suggests the survival impact of *MYC* mRNA expression is still strong even without the effect of *BCL2,* that *MYC* mRNA is an independent impact factor (Figure 5-10 bottom right and Table 5-7).

Table 5-7: Survival difference between various groups: showed in p-values

	CHOP		**RCHOP**	
	GSE4732	**GSE10846**	**GSE10846**	**GSE31312**
MYC and BCL2 co-expression: both-high, myc-high-bcl2-low,myc-low-bcl2-high,both-low				
All subtypes	0.004	<0.001	0.014	0.45
GCB	0.03	0.01	0.159	0.813
ABC	0.509	0.491	0.692	0.088
MYC expression 2 categories: high and low				
in BCL2 low	0.13	0.231	0.066	0.458
in BCL2 high	0.001	0.006	0.034	0.161
Multivariate cox proportional hazard model				
Age	0.002	<0.001	0.01	No age info
Subtype	0.371	0.003	<0.001	<0.001
Predict *myc-r*	0.658	0.129	0.03	0.09
MYC expression	0.02	0.06	0.01	0.08
BCL2 expression	0.15	0.111	0.95	0.74

In addition, we performed a more sophisticated multiple covariate analysis by the Cox proportional hazard model, and the results are presented in Table 5-7. It shows that age and subtypes are relatively consistent impact factors (significant in over three datasets), and that *MYC* expression level also correlates to survival although at the edge of significance in two datasets. *BCL2* expression doesn't seem to correlate with survival obviously in any dataset.

5.3 *MYC* potential mechanism exploration

MYC mechanism in non-Burkitt lymphomas remains elusive despite extensive investigations. In this section we explore this topic first by performing gene set enrichment analysis on biologically distinct groups. We conducted GSEA on the *MYC* dataset (combination of three public datasets) used in section 5.2.2, as well as the samples that have very high expression against very low expression of *MYC* mRNA, because the two groups present a different clinical course according to the results in section 5.3.

Then we investigated the potential *MYC* activity by identifying genes that are associated with *MYC*, and explore the biological meanings of the genes with DAVID function analysing tool. The *MYC*-associated gene lists were selected in three different ways: (1) differentially expressed genes between FISH detected *MYC*-rearranged and *MYC*-negative GCB subtypes, since above analysis (section 5.3.1) shows that *MYC* rearrangement has a significant impact in GCB cases; (2) differentially expressed between very-high against very-low *MYC* mRNA expression level groups, and (3) genes have strong positive/negative correlation with *MYC* mRNA expression.

In addition to GSE10846 and GSE31312 datasets applied in the survival analysis, another four public datasets GSE12195 [196], GSE22470 [192], GSE34171 [197] and Monti data [198] were used to select *MYC*-associated genes and perform functional discovery. The detail is listed in Table 5-8.

Table 5-8: Additional datasets used in mechanism exploration

Dataset	Description	Tissue[1]	Platform
GSE10846	414 DLBCL	FF	Affymetrix U133plus2
GSE31312	498 DLBCL	FFPE	Affymetrix U133plus2
GSE12195	73 DLBCL	FF	Affymetrix U133plus2
GSE22470	271DLBCL	FF	Affymetrix U133a
GSE34171	91DLBCL	FF	Affymetrix U133plus2
Monti	176DLBCL	FF	Custom array

[1]FF denotes fresh frozen, and FFPE denotes formalin fixed paraffin embedded tissue

5.3.1. Gene set enrichment analysis

We use the GSEA software that developed by the Broad Institute and the annotated dataset collection molecular signature database (MisgDB)[178] to assess what datasets are enriched in our expression data. There are 7 collections (c1 to c7) of datasets in the MisgDB 4.0, and here we included c2 (curated gene sets from online pathway databases, publications in PubMed, and knowledge of domain experts.), c5 (GO gene sets consist of genes annotated by the same GO terms), c6 (oncogenic signatures defined directly from microarray gene expression data from cancer gene perturbations), and c7 (immunologic signatures defined directly from microarray gene expression data from immunologic studies). And we exclude the datasets which have a number of genes less than 15 or over 200.

In the *MYC* dataset (a combination of three public datasets, see section 5.2.2), an analysis between *MYC*-rearranged and *MYC*-negative samples, there are 11 gene sets are significantly (p-value <0.01) enriched in rearranged phenotype (see Figure 5-13 and details for Appendix B table 1), including the genes (up-regulated in BL) applied in the molecular BL classifier generated by Hummel group:

HUMMEL_BURKITTS_LYMPHOMA_UP

And of genes strongly up-regulated in B493-6 cells (B lymphocytes) by a combination of MYC and serum but not by each of them alone:

SCHLOSSER_*MYC*_AND_SERUM_RESPONSE_SYNERGY

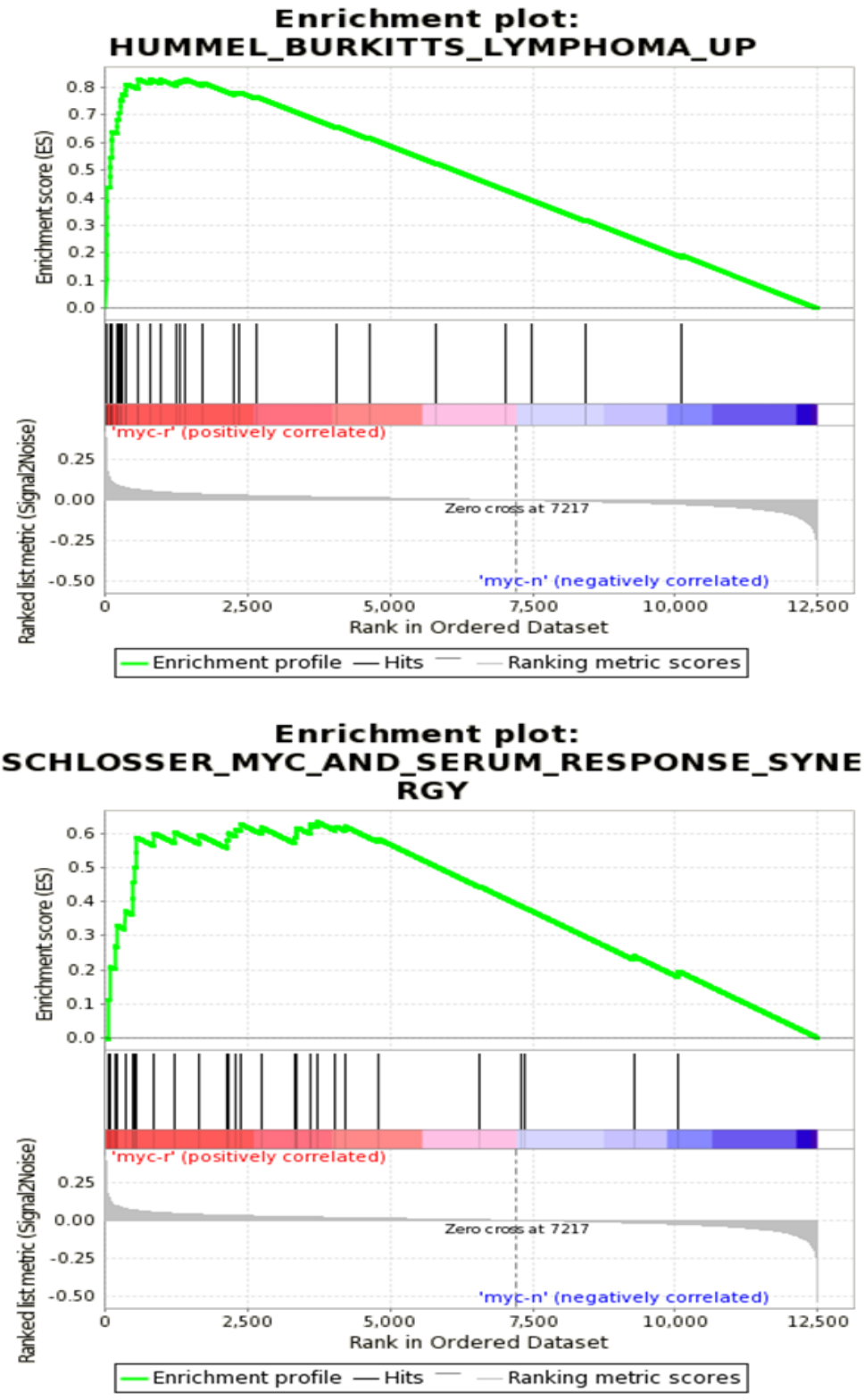


Figure 5-12: Enriched gene sets in *MYC*-rearranged groups.

The gene sets came out suggests there is an association between *MYC* translocation and other MYC activities. However a problem is none of the enriched gene sets has a false discover rate (FDR) less than 0.05, which indicates the significance drop down when comparing with other gene sets in the empirical null distribution. This could be because we do not have enough samples. Next we performed the same analysis on the expression profiles of samples that have the top 20 percent and bottom 20 percent of *MYC* expression level in each of the six data sets listed in Table 5-8.

The number of enriched gene sets picked out in each dataset differ vastly, there are 195 gene set in Monti dataset, 49 gene set in GSE10846 and one gene set in GSE34171 that have normalized enrichment score (NES) over 2.0 and FDR less than 0.05, while no gene set was chosen as enriched in the remaining three datasets (full results see Appendix B table 2-3). Again this could because of the small number of the samples in the analysis (15 in each group of GSE12195, 53 in GSE22470), and this could also be because there is less biological information in FFPE samples (GSE31312). However some of the enriched gene sets selected by the analysis were interesting, a few examples are listed below:

MYC_UP.V1_UP,

which are genes up-regulated in primary epithelial breast cancer cell culture over-expressing MYC gene, and genes up-regulated by MYC according to the MYC Target Gene Database (see Figure 5-14, for example):

DANG_REGULATED_BY_MYC_UP

Also there are some reactome gene sets related to various activities such as:

REACTOME_TRANSLATION

REACTOME_INFLUENZA_LIFE_CYCLE

REACTOME_NONSENSE_MEDIATED_DECAY_ENHANCED_BY_THE_EXON_JUNCTION_COMPLEX

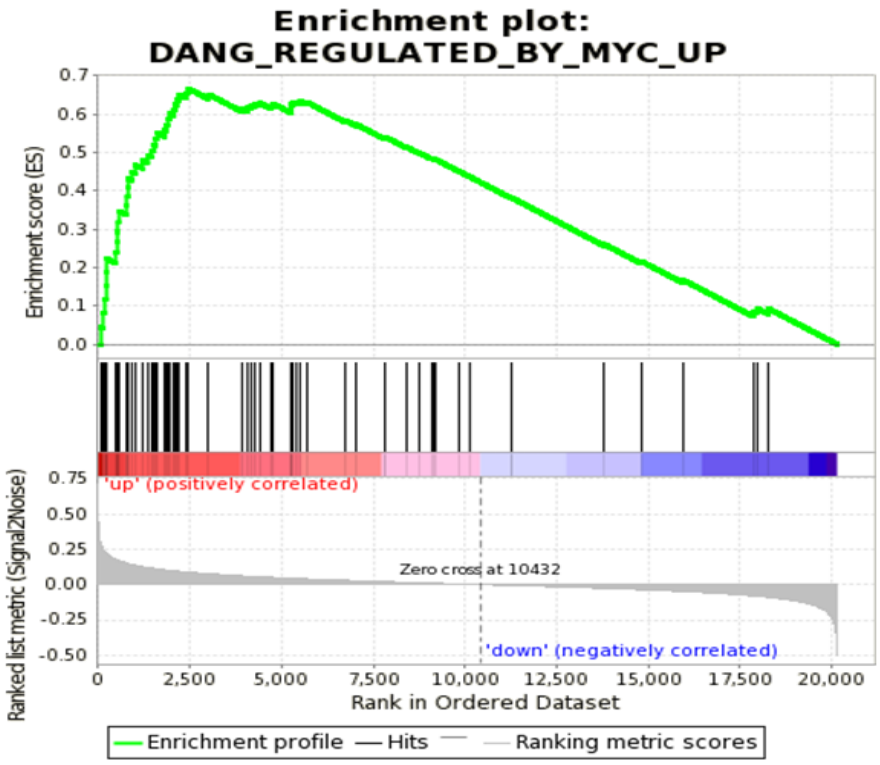


Figure 5-13: Enriched gene sets in top 20 percent *MYC* high expressed phenotype.

5.3.2. Select *MYC*-associated gene lists

The differentially expressed genes between the *MYC*-rearranged GCB cases were selected in limma package with the same criteria as in section 5.2.2, and there are 40 significantly up-regulated plus 60 down-regulated genes respectively. Next we explain how the other two lists of genes are selected.

First are the genes that are different between *MYC* very high (top 20%) and very low (bottom 20%) groups. We selected the genes in each of above six datasets using the limma package by setting the criteria at p-value less than 0.01 and log fold change over 0.5, and the genes that are picked out by at least 4 dataset are called significantly differentially expressed. There are 137 genes came out as significantly up-regulated in *MYC* very high cases, and only 3 genes as significantly down-regulated.

Second gene lists are selected according to the *MYC* expression correlation: all samples in each of the six datasets are re-classified into ABC, GCB, and UCL subgroups by a DLBCL subtype GEP classifier developed by collaborator [199]. For each subtype, all genes are ranked by an adjusted spearman correlation with *MYC* expression, which is the average correlation across all datasets multiplied by the number of datasets that the gene appears in. The average and the standard deviation of the ranked genes among all subtypes are showed in Figure 5-12. The genes that have large rank are positive-correlated with *MYC* expression, and down in the rank list are negative-correlated with *MYC* expression.

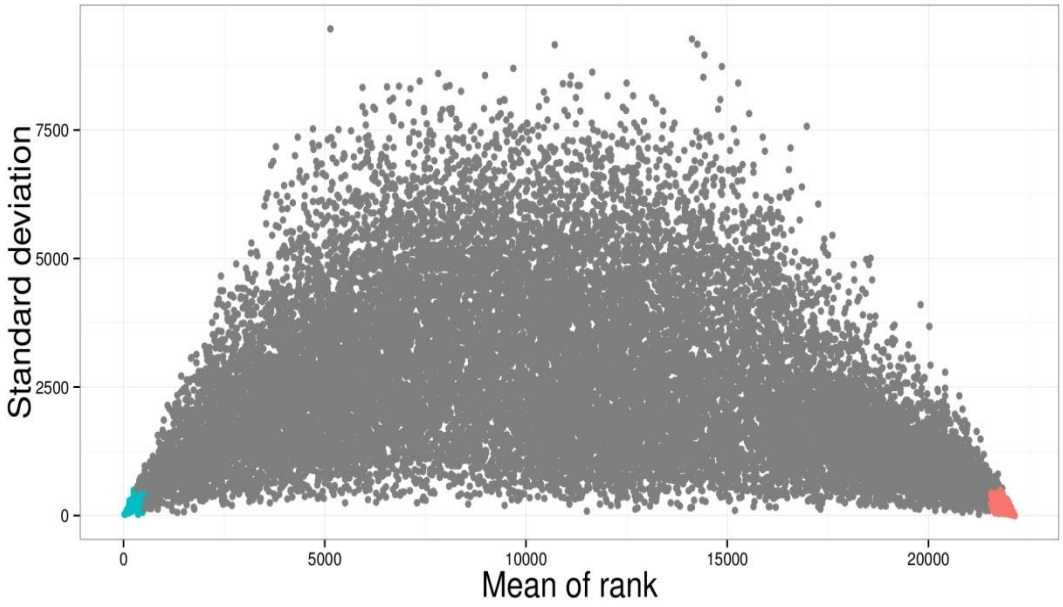


Figure 5-14: Mean and standard deviation of the *MYC* expression correlation gene rank lists for all subtypes.

The median of the standard deviation for gene ranks in different subtypes is 2221, which is about 10% of all genes. And it seems that the genes have high positive correlation or negative correlation (spearman correlation) with *MYC* expression is quite consistent, both top and bottom of the average rank list have relatively small deviation. Here we chose the top 200 plus bottom 200 genes in the rank list which also have standard deviation smaller than

100 to perform further analysis. And there are 90 positive correlated and 12 negative correlated genes selected respectively.

5.3.3. DAVID functional analysis

We use DAVID bioinformatics resources to explore the potential biological meanings of *MYC*-associated gene lists. Gene function was annotated on GOTERM_BP_ALL (gene ontology terms of biological process), SP_PIR_KEYWORDS (Swiss-Prot and Protein Information Resource), and KEGG_PATHWAY for gene ontology, function network and canonical pathway analysis respectively.

For the *MYC* rearrangement signature gene list in GCB subtypes, no significant (Benjamini p-value < 0.01 and FDR < 0.05) annotations were selected as enriched in the up-regulated gene set, while for the down-regulated genes, the GO term annotations related to regulation of apoptosis or cell death stand out significantly (see Table 5-9).

Table 5-9: GO term annotation of down-regulated genes in *MYC*-rearranged GCB

Category	%	p-value	Benjamini	FDR
regulation of apoptosis	24.07	1.07-E05	9.75-E03	0.02
regulation of programmed cell death	24.07	1.18-E05	5.40-E03	0.02
regulation of cell death	24.07	1.22-E05	3.74-E03	0.02

For the genes selected between *MYC* very high against very low mRNA expression groups, there are few genes selected as down-regulated in the *MYC* very high expressed group, and we analysed the up-regulated gene list. The annotation for GO term, functions and pathways are listed in 5-11. It shows that the main functions of the genes are proliferating and metabolic related such as ribosome biogenesis, RNA processing, DNA replication, and metabolic process of rRNA or nuclearRNAs.

Table 5-10: DAVID annotation for up-regulated genes in *MYC* highly expressed cases

Category	%	p-value	Benjamini	FDR
GOTERM_BP_ALL				
ribonucleoprotein complex biogenesis	18.80	1.29E-21	1.13E-18	2.00E-18
ribosome biogenesis	15.79	5.69E-20	2.51E-17	8.85E-17
rRNA processing	9.77	3.64E-11	1.07E-08	5.65E-08
rRNA metabolic process	9.77	6.07E-11	1.34E-08	9.44E-08
ncRNA metabolic process	12.78	2.73E-10	4.82E-08	4.25E-07
RNA processing	18.05	6.55E-10	9.62E-08	1.02E-06
ncRNA processing	11.28	1.48E-09	1.86E-07	2.30E-06
DNA metabolic process	17.00	1.28E+01	0.00	0.02
SP_PIR_ KEYWORDS				
acetylation	61.654	3.36E-37	7.02E-35	4.24E-34
ribosome biogenesis	9.774	3.64E-16	3.47E-14	4.22E-13
nucleus	50.376	2.28E-12	1.59E-10	2.87E-09
phosphoprotein	67.669	4.19E-12	2.19E-10	5.29E-09
rrna processing	6.015	1.96E-07	8.18E-06	0.000
mitochondrion	16.541	2.33E-07	8.13E-06	0.000
atp-binding	21.053	2.77E-07	8.26E-06	0.000
nucleotide-binding	22.556	2.96E-06	7.72E-05	0.004
Chaperone	6.767	1.35E-05	0.000	0.017
transit peptide	10.526	2.46E-05	0.001	0.031
dna replication	5.263	3.19E-05	0.001	0.040
KEGG_PATHWAY				
Purine metabolism	8.271	2.42E-06	0.000	0.002
Cysteine and methionine metabolism	4.511	2.04E-05	0.001	0.021

For the genes correlated with *MYC* expression, again nothing was significant for the negative-correlated genes, and the results for positive-correlated genes are shown in Table 5-12. Similar to the genes up-regulated in *MYC* very highly expressed cases; the positive-correlated genes have predominant functions involving ribosome biogenesis and protein biosynthesis.

Table 5-11: DAVID annotation for *MYC* positive-correlated genes

Category	%	p-value	Benjamini	FDR
Gene ontology				
ribonucleoprotein complex biogenesis	1.050	3.69E-10	3.19E-07	5.73E-07
ribosome biogenesis	0.808	2.82E-08	1.22E-05	4.38E-05
Function				
acetylation	5.250	2.05E-37	3.51E-35	2.51E-34
phosphoprotein	4.927	1.65E-09	1.41E-07	2.01E-06
ribosome biogenesis	0.565	7.07E-08	4.03E-06	8.63E-05
cytoplasm	2.989	7.89E-08	3.37E-06	9.64E-05
nucleotide-binding	2.019	2.13E-07	0.000	0.000
atp-binding	1.777	2.62E-07	7.46E-06	0.000
Chaperone	0.727	5.35E-07	1.31E-05	6.53E-04
host-virus interaction	0.889	5.72E-07	1.22E-05	6.99E-04
mitochondrion	1.373	7.35E-07	1.40E-05	8.98E-04
ribonucleoprotein	0.808	4.33E-06	7.40E-05	5.29E-03
nucleus	3.150	4.93E-06	7.66E-05	0.006
protein biosynthesis	0.646	2.13E-05	3.04E-04	0.026

5.3.4. Potential PRMT5 involvement

An interesting finding of the gene lists that are positive-correlated with *MYC* expression is that *PRMT5* appears at a very high rank in all subtypes. The PRMT5 protein (protein arginine N-methyltransferase 5) is the major enzyme

that is responsible for mono- and symmetric dimethylation of arginine, whereas few literatures have reported the connection between MYC and PRMT5, only a study published in 2003 showed that PRMT5 and mSin3A interact with the same hSWI/SNF (nucleosome remodelling complex, which is a group of proteins that associate to remodel the way DNA is packaged) subunits as those targeted by MYC [200], and that PRMT5 is directly recruited to MYC target gene *CAD* promoter.

However, an expanding literature have demonstrated its critical biological function in a wide range of cellular processes including histone methylation, genome organization, chromatin regulation, RNA processing, proliferation and more [201-204]. More recently the role of PRMT5 in lymphoma has becoming highly appreciated, it is strongly suggested that PRMT5 is required for lymphoma genesis and the overexpression is found to be involved in the proliferation and survival of mantle cell (MCL) and DLBCL cells [205-207], moreover a few studies have shown that a small molecule inhibitor of PRMT5 could kill lymphoma cells [34, 205], and a validation of PRMT5 as a candidate therapeutic target in glioblastoma in a mouse model has successfully enhanced the cell survival [208], which are all promising in offering a therapeutic strategy for lymphoma patients.

5.4 Conclusion

In this chapter we investigated a few rather keen topics related to *MYC*-associated non-Burkitt lymphomas. First we tried to differentiate the FISH detected *MYC*-rearranged cases by finding a characteristic expression pattern. Although the expression pattern was able to correctly identify most of the rearranged samples, it doesn't seem to have full concordance with the FISH detected status in non-BL comparing to BL samples. This may because *MYC*-rearranged non-BL is often accompanied with various types of aberrations and complex karyotypes, and it is difficult to find an expression pattern that effectively represents this heterogeneous group. However the significant lower *MYC* single gene expression of the predicted *myc-n* group compared to *myc-r* group in FISH detected *MYC*-rearranged

cases may suggest that *MYC*-rearrangement do not necessarily lead to high-level expression of *MYC* mRNA.

The survival analysis of the *MYC* rearrangement expression pattern showed that it had significant worse impact on GCB subtype but not in other subtypes. More closely analysis on *MYC* and *BCL2* mRNA expression revealed that high *MYC* expression is strongly correlated with short survival, while the inferior effect of *BCL2* expression decreases when the treatment improved to R-CHOP. It's no doubt that certain obstacles exist in this type of survival analysis, especially dataset GSE31312 was generated from FFPE samples, and that it doesn't seem to give any good correlation between survival and single gene expression. Still high level of *MYC* mRNA expression showed rather confident evidence of affecting patient's outcome, even in consideration with age and molecular subtypes prognostic factors.

The data presented here suggest two potentially important conclusions. First, the analysis of our DASL data suggests that a *MYC* translocation identified by FISH does not always lead to a high level of *MYC* expression and the expression of the overall gene expression signature identified in many re-arranged cases. Second, that *MYC* expression levels are the best prognostic indicator compared with *MYC* translocations for many DLBCLs. Putting these two together you can make the suggestion that *MYC* expression levels are the best prognostic marker to apply in these cases.

At last we explored the potential involved mechanism of MYC in non-BLs. The *MYC* rearrangement expression pattern genes seem to have a relation with down-regulating cell apoptosis. And the *MYC* expression correlated genes have enriched functions like RNA/ribosome activity and DNA replication. In addition, a potential treatment target PRMT5 is found to be highly correlated with *MYC* expression.

Chapter 6
Discussion and future work

6.1 Discussion

The work carried out in this project presents a systematic investigation of problems proposed in the introduction chapter, regarding to developing a gene expression profiling based Burkitt lymphoma and diffuse large B-cell lymphoma classifier that is able to work effectively on formalin-fixed paraffin-embedded samples commonly used in routine clinics. Another aim of the study is to explore the role of *MYC* in the non-Burkitt B-cell lymphomas that raised great attention recently due to its particularly aggressive clinical course and lack of sufficient treatment paradigms. A summary of the results from each chapter is listed below:

In chapter three, we have shown that two previous different classifiers in the research literature can be well recapitulated with the LibSVM classifier constructed by a much smaller gene set than those used in original classifiers. And that the classifier trained on the previous two corresponding datasets can be successfully transferred to other public datasets with similar BL definition as long as the cross-normalization is performed. However the classification results reflect that a less strict BL definition than which applied in previous classifiers is needed, in order to recognize the BLs from other datasets including some of which may have a weaker signal.

Dependence on training data highlights the underlying difficulty in this and many similar studies, which is the lack of a ‘gold standard’ against which to evaluate new classifiers. Even though disease categories like BL and DLBCL have developed over many years with a variety of phenotypic and molecular diagnostic criteria, there are still a significant number of cases that are complex and neither expert pathological assessors nor recent molecular classifiers can effectively distinguish them. And this becomes even more obvious when trying to combine studies by different research groups.

Therefore an alternative evaluation that can be performed is to combine the analysis of survival separation and treatment response.

In chapter four, we validated the classifier on the FFPE datasets produced by our collaborators. The classification results of samples performed on two platforms show a generally good reproducibility as well as high consistency among cross-platform normalization methods. Nevertheless, the classifiers trained with the BL definition applied in previous literatures are less sensitive in recognizing BLs with weaker signals, which is more likely to happen in FFPE samples. And the classifier trained with an adjusted training set assigns samples into BL and DLBCL categories highly concordant with the diagnosis by the clinicians.

More importantly, the examination of the outcome on the RCHOP treated *MYC* rearranged DLBCL cases shows that, the cases classified as BL have worse response than those classified as DLBCL, indicating the classifier is of potential clinical importance. Whereas the number of cases analyzed in treatment response is rather limited to draw statistically powerful conclusions yet, hence a next step of the work is to test if this treatment separation by our classifier still stands as we gather more data. However, it is important to note that this is one of the typical issues because even with a large data set such as this, when the cases are stratified by many clinical features or that the level of agreement between diagnosis is high there will only be small number cases of interest left to examine, particularly, the treatment options in the setting of B-cell malignancies usually improves at a high rate, thus the use of clinical outcome with a specific therapy could become an unstable parameter and add more stratification of the evaluation.

Overall, our study of the discrimination of BL and DLBCL thoroughly performed the comparison of previous GEP classifiers, different platforms, normalization tools and datasets from various sample cohorts. The results show an excellent transferability and reproducibility among different research groups as well as FFPE samples from the clinical archive, which is a valuable practical step towards the refinement of pathological diagnosis and therapeutic approaches. The new classifier proposed here has the obvious stronger robustness comparing to previous research GEP classifiers that

developed and validated based on a single dataset/platform. And it is more reliable and objective than the classification based on immunohistochemistry method, which highly rely on the opinion of the technicians. Further, with new technologies on the way, such as wider mutation profiling, the technique is not yet outcompeted and the classifier can also provide as an external validation from the GEP angle. In addition, the study might also be adapted to some novel techniques such as NanoString and QuantiGene that are likely to be use in a routine diagnostic approach.

In chapter five, we explored the *MYC* rearrangement expression pattern, the clinical impact of *MYC* on non-Burkitt B-cell lymphoma as well as the potential mechanism underlying MYC deregulation. The results show that most *MYC* rearranged cases express a particular pattern, yet the expression pattern do not always present in the FISH detected rearranged cases. It also shows that *MYC* mRNA high expression level is significantly correlated with poor outcome both in CHOP and RCHOP treated cases, and is a stronger survival indicator in non-Burkitt lymphomas comparing to *MYC* translocation or *MYC* translocation expression pattern, and *BCL2* mRNA expression. As far as the suggests that the protein co-expression of high level MYC and BCL2 cause the basis for the inferior outcome of non-Burkitt lymphomas, the high-level co-expression of the mRNA expression also show obvious worse survival, however, the impact of *MYC* mRNA high expression alone is not fully justified in the R-CHOP treated dataset. This indicates high level of *MYC* mRNA expression serves as an independent prognostic impact factor, and a future step is to validate a reproducible way that can separate these cases.

The gene set enrichment analysis between the samples expressing high level and low level of *MYC* mRNA identified enriched gene sets that consist of genes involved in other cancer cells showing MYC overexpression, and gene sets from MYC target gene database. The functional analysis of *MYC* high mRNA expression related genes and *MYC* mRNA expression correlated genes show a predominant functions relating to proliferation and metabolic activities. A positive finding is that *PRMT5* is strongly co-expressed with *MYC* irrespective of DLBCL subtypes, which suggest PRMT5 is potentially involved in the biological mechanism in *MYC*-

associated non-Burkitt lymphomas. In fact, PRMT5 overexpression is found to be related with the proliferation and survival of DLBCL cells, and recently a few studies have shown that a small molecule inhibitor of PRMT5 could kill lymphoma cells, which provides a possible treatment direction.

The analysis of *MYC*-associated non-Burkitt lymphoma is largely restricted by the fact that not enough data is available currently; in addition, it is complicated by the massive heterogeneity introduced by the complex phenotype and molecular features, the various *MYC* aberrancy detection methods with none widely accepted as standard, even in survival analysis the treatment schemes are sometimes various patient by patient and when combining data from multiple centers, other inconsistencies such as the diagnostic criteria could also be introduced. Therefore, typical difficulty of the related studies is that they usually lead to rather small sample set and weak conclusions. Given the current situation, there is a clear need to uniformly define the *MYC*-associated non-Burkitt lymphoma, and design specific clinical trials for these cases to gather more samples over time.

6.2 Future work

This study confirms that earlier work on gene expression based definitions of BL and DLBCL can be adapted for routine use to produce an automatic classifier with a high degree of concordance with more traditional methods, however the lack of clarity of the intermediate cases still exists on the molecular level with the treatment decisions remaining difficult. Moreover, new findings such as BL are associated with distinct mutation spectra from those observed in DLBCL, including *ID3* and *MYC*, and that there are also genes mutated in both DLBCL and BL with similar frequencies such as *GNA13* all suggests that evaluating these molecular features is more than gene expression but to combine a wider mutation profiling. It is likely that a combination of both information sources as the basis of future classifiers could lead to increased robustness in the context of heterogeneous diseases and the inevitable noise associated with all measurements on clinical samples.

Another approach related to BL and DLBCL diagnosis and treatment decision worth trying out in the future is that, instead of classifying samples into different categories or the stratifying them based on various features, which commonly lead to smaller sample set, we could perform a similarity query of a specific patient according to its gene expression profiles, maybe in addition with phenotype, clinical features as well as other genetic information. The method would then return a list of patients that have the overall high similarity with the query patient, which can be used to assist the diagnosis and predict the prognosis on a particular treatment. The advantage underlying this is that, not like classifiers which sometimes assign an intermediate case to a class with less confidence, it finds the patients most close to the case to make sure the query is most informative.

This study also suggests that high level of *MYC* mRNA expression acts as a stable prognostic factor and is potentially able to represent *MYC* aberrations. Moreover, there is strong evidence that *PRMT5* is correlated with *MYC* high expression which opens a potential therapeutic target for *MYC*-associated non-Burkitt lymphomas. PRMT5 over expression is known to be involved in the proliferation and survival of DLBCLs, and recent studies have suggested that the inhibition of PRMT5 could offer a promising therapeutic strategy for lymphoma patients [205, 209]. Therefore, the next step following this finding would be to test whether interfering the PRMT5 activity in the *MYC* high expressed B-cell lymphoma cells could lead to the cell death.

List of References

1. S.H.Swerdlow EC, N.L. Harris, et al. (ed.): **WHO classification of tumours of haematopoietic and lymphoid tissues, Fourth Edition**. Lyon, France: IARC Press; 2008.
2. Fuller GN: **The WHO classification of tumours of the central nervous system, 4th edition**. *Arch Pathol Lab Med* 2008, **132**(6):906-906.
3. Bociek RG: **Adult Burkitt's lymphoma**. *Clin Lymphoma* 2005, **6**(1):11-20.
4. Molyneux EM, Rochford R, Griffin B, Newton R, Jackson G, Menon G, Harrison CJ, Israels T, Bailey S: **Burkitt's lymphoma**. *Lancet* 2012, **379**(9822):1234-1244.
5. Molyneux EM, Rochford R, Griffin B, Newton R, Jackson G, Menon G, Harrison CJ, Israels T, Bailey S: **Burkitt's lymphoma**. *Lancet* 2012, **379**(9822):1234-1244.
6. Fujita S, Buziba N, Kumatori A, Senba M, Yamaguchi A, Toriyama K: **Early stage of Epstein-Barr virus lytic infection leading to the "starry sky" pattern formation in endemic Burkitt lymphoma**. *Arch Pathol Lab Med* 2004, **128**(5):549-552.
7. Hecht JL, Aster JC: **Molecular biology of Burkitt's lymphoma**. *J Clin Oncol* 2000, **18**(21):3707-3721.
8. Love C, Sun Z, Jima D, Li G, Zhang J, Miles R, Richards KL, Dunphy CH, Choi WW, Srivastava G *et al*: **The genetic landscape of mutations in Burkitt lymphoma**. *Nat Genet* 2012, **44**(12):1321-1325.
9. Mead GM, Sydes MR, Walewski J, Grigg A, Hatton CS, Pescosta N, Guarnaccia C, Lewis MS, McKendrick J, Stenning SP *et al*: **An international evaluation of CODOX-M and CODOX-M alternating with IVAC in adult Burkitt's lymphoma: results of United Kingdom Lymphoma Group LY06 study**. *Ann Oncol* 2002, **13**(8):1264-1274.
10. Mead GM, Barrans SL, Qian W, Walewski J, Radford JA, Wolf M, Clawson SM, Stenning SP, Yule CL, Jack AS: **A prospective clinicopathologic study of dose-modified CODOX-M/IVAC in patients with sporadic Burkitt lymphoma defined using cytogenetic and immunophenotypic criteria (MRC/NCRI LY10 trial)**. *Blood* 2008, **112**(6):2248-2260.
11. Martelli M, Ferreri AJ, Agostinelli C, Di Rocco A, Pfreundschuh M, Pileri SA: **Diffuse large B-cell lymphoma**. *Crit Rev Oncol Hematol* 2013, **87**(2):146-171.
12. Sehn LH, Donaldson J, Chhanabhai M, Fitzgerald C, Gill K, Klasa R, MacPherson N, O'Reilly S, Spinelli JJ, Sutherland J *et al*: **Introduction of combined CHOP plus rituximab therapy dramatically improved outcome of diffuse large B-cell lymphoma in British Columbia**. *J Clin Oncol* 2005, **23**(22):5027-5033.

13. Rosenwald A, Wright G, Chan WC, Connors JM, Campo E, Fisher RI, Gascoyne RD, Muller-Hermelink HK, Smeland EB, Giltnane JM *et al*: **The use of molecular profiling to predict survival after chemotherapy for diffuse large-B-cell lymphoma**. *N Engl J Med* 2002, **346**(25):1937-1947.
14. Rimsza LM, LEBLanc ML, Unger JM, Miller TP, Grogan TM, Persky DO, Martel RR, Sabalos CM, Seligmann B, Braziel RM *et al*: **Gene expression predicts overall survival in paraffin-embedded tissues of diffuse large B-cell lymphoma treated with R-CHOP**. *Blood* 2008, **112**(8):3425-3433.
15. Gutierrez-Garcia G, Cardesa-Salzmann T, Climent F, Gonzalez-Barca E, Mercadal S, Mate JL, Sancho JM, Arenillas L, Serrano S, Escoda L *et al*: **Gene-expression profiling and not immunophenotypic algorithms predicts prognosis in patients with diffuse large B-cell lymphoma treated with immunochemotherapy**. *Blood* 2011, **117**(18):4836-4843.
16. Barrans SL, Crouch S, Care MA, Worrillow L, Smith A, Patmore R, Westhead DR, Tooze R, Roman E, Jack AS: **Whole genome expression profiling based on paraffin embedded tissue can be used to classify diffuse large B-cell lymphoma and predict clinical outcome**. *Br J Haematol* 2012, **159**(4):441-453.
17. Coiffier B: **State-of-the-art therapeutics: diffuse large B-cell lymphoma**. *J Clin Oncol* 2005, **23**(26):6387-6393.
18. Haralambieva E, Boerma EJ, van Imhoff GW, Rosati S, Schuuring E, Muller-Hermelink HK, Kluin PM, Ott G: **Clinical, immunophenotypic, and genetic analysis of adult lymphomas with morphologic features of Burkitt lymphoma**. *Am J Surg Pathol* 2005, **29**(8):1086-1094.
19. Bellan C, Stefano L, Giulia de F, Rogena EA, Lorenzo L: **Burkitt lymphoma versus diffuse large B-cell lymphoma: a practical approach**. *Hematol Oncol* 2010, **28**(2):53-56.
20. Nakamura N, Nakamine H, Tamaru J, Nakamura S, Yoshino T, Ohshima K, Abe M: **The distinction between Burkitt lymphoma and diffuse large B-Cell lymphoma with c-myc rearrangement**. *Mod Pathol* 2002, **15**(7):771-776.
21. Nomura Y, Karube K, Suzuki R, Ying G, Takeshita M, Hirose S, Nakamura S, Yoshino T, Kikuchi M, Ohshima K: **High-grade mature B-cell lymphoma with Burkitt-like morphology: results of a clinicopathological study of 72 Japanese patients**. *Cancer Sci* 2008, **99**(2):246-252.
22. de Jong D: **Novel lymphoid neoplasms--the borderland between diffuse large B-cell lymphoma and Burkitt's lymphoma**. *Haematologica* 2009, **94**(7):894-896.
23. Bellone M, Zaslav AL, Ahmed T, Lee HL, Ma Y, Hu Y: **IGH amplification in patients with B cell lymphoma unclassifiable, with features intermediate between diffuse large B cell lymphoma and Burkitt's lymphoma**. *Biomark Res* 2014, **2**:9.
24. Thomas DA, O'Brien S, Faderl S, Manning JT, Jr., Romaguera J, Fayad L, Hagemeister F, Medeiros J, Cortes J, Kantarjian H: **Burkitt lymphoma and atypical Burkitt or Burkitt-like lymphoma: should**

these be treated as different diseases? *Curr Hematol Malig Rep* 2011, **6**(1):58-66.

25. Burgesser MV, Gualco G, Diller A, Natkunam Y, Bacchi CE: **Clinicopathological features of aggressive B-cell lymphomas including B-cell lymphoma, unclassifiable, with features intermediate between diffuse large B-cell and Burkitt lymphomas: a study of 44 patients from Argentina**. *Ann Diagn Pathol* 2013, **17**(3):250-255.
26. Salaverria I, Siebert R: **The Gray Zone Between Burkitt's Lymphoma and Diffuse Large B-Cell Lymphoma From a Genetics Perspective**. *Journal of Clinical Oncology* 2011, **29**(14):1835-1843.
27. Mossafa H, Damotte D, Jenabian A, Delarue R, Vincenneau A, Amouroux I, Jeandel R, Khoury E, Martelli JM, Samson T *et al*: **Non-Hodgkin's lymphomas with Burkitt-like cells are associated with c-Myc amplification and poor prognosis**. *Leuk Lymphoma* 2006, **47**(9):1885-1893.
28. Eilers M, Eisenman RN: **Myc's broad reach**. *Gene Dev* 2008, **22**(20):2755-2766.
29. Dang CV: **MYC on the Path to Cancer**. *Cell* 2012, **149**(1):22-35.
30. Ouansafi I, He B, Fraser C, Nie K, Mathew S, Bhanji R, Hoda R, Arabadjief M, Knowles D, Cerutti A *et al*: **Transformation of follicular lymphoma to plasmablastic lymphoma with c-myc gene rearrangement**. *Am J Clin Pathol* 2010, **134**(6):972-981.
31. Meyer N, Penn LZ: **Reflecting on 25 years with MYC**. *Nat Rev Cancer* 2008, **8**(12):976-990.
32. Adhikary S, Eilers M: **Transcriptional regulation and transformation by Myc proteins**. *Nat Rev Mol Cell Biol* 2005, **6**(8):635-645.
33. Luscher B, Vervoorts J: **Regulation of gene transcription by the oncoprotein MYC**. *Gene* 2012, **494**(2):145-160.
34. Sabo A, Kress TR, Pelizzola M, de Pretis S, Gorski MM, Tesi A, Morelli MJ, Bora P, Doni M, Verrecchia A *et al*: **Selective transcriptional regulation by Myc in cellular growth control and lymphomagenesis**. *Nature* 2014, **511**(7510):488-492.
35. Nie Z, Hu G, Wei G, Cui K, Yamane A, Resch W, Wang R, Green DR, Tessarollo L, Casellas R *et al*: **c-Myc is a universal amplifier of expressed genes in lymphocytes and embryonic stem cells**. *Cell* 2012, **151**(1):68-79.
36. Lin CY, Loven J, Rahl PB, Paranal RM, Burge CB, Bradner JE, Lee TI, Young RA: **Transcriptional amplification in tumor cells with elevated c-Myc**. *Cell* 2012, **151**(1):56-67.
37. Ott G, Rosenwald A, Campo E: **Understanding MYC-driven aggressive B-cell lymphomas: pathogenesis and classification**. *Blood* 2013, **122**(24):3884-3891.
38. Klapproth K, Wirth T: **Advances in the understanding of MYC-induced lymphomagenesis**. *Br J Haematol* 2010, **149**(4):484-497.
39. Li S, Seegmiller A, Lin P, Medeiros LJ: **High-Grade B-Cell Lymphomas with Concurrent MYC and BCL2 Abnormalities Other Than Translocations Behave Similarly to MYC/BCL2 Double Hit Lymphomas**. *Lab Invest* 2013, **93**:341a-341a.

40. Li S, Lin P, Young KH, Kanagal-Shamanna R, Yin CC, Medeiros LJ: **MYC/BCL2 double-hit high-grade B-cell lymphoma**. *Adv Anat Pathol* 2013, **20**(5):315-326.
41. Petrich AM, Nabhan C, Smith SM: **MYC-associated and double-hit lymphomas: a review of pathobiology, prognosis, and therapeutic approaches**. *Cancer* 2014, **120**(24):3884-3895.
42. Le Gouill S, Talmant P, Touzeau C, Moreau A, Garand R, Juge-Morineau N, Gaillard F, Gastinne T, Milpied N, Moreau P *et al*: **The clinical presentation and prognosis of diffuse large B-cell lymphoma with t(14;18) and 8q24/c-MYC rearrangement**. *Haematologica* 2007, **92**(10):1335-1342.
43. Johnson NA, Savage KJ, Ludkovski O, Ben-Neriah S, Woods R, Steidl C, Dyer MJ, Siebert R, Kuruvilla J, Klasa R *et al*: **Lymphomas with concurrent BCL2 and MYC translocations: the critical factors associated with survival**. *Blood* 2009, **114**(11):2273-2279.
44. Akyurek N, Uner A, Benekli M, Barista I: **Prognostic significance of MYC, BCL2, and BCL6 rearrangements in patients with diffuse large B-cell lymphoma treated with cyclophosphamide, doxorubicin, vincristine, and prednisone plus rituximab**. *Cancer* 2012, **118**(17):4173-4183.
45. Aukema SM, Kreuz M, Kohler CW, Rosolowski M, Hasenclever D, Hummel M, Kuppers R, Lenze D, Ott G, Pott C *et al*: **Biological characterization of adult MYC-translocation-positive mature B-cell lymphomas other than molecular Burkitt lymphoma**. *Haematologica* 2014, **99**(4):726-735.
46. Aukema SM, Siebert R, Schuuring E, van Imhoff GW, Kluin-Nelemans HC, Boerma EJ, Kluin PM: **Double-hit B-cell lymphomas**. *Blood* 2011, **117**(8):2319-2331.
47. Niitsu N, Okamoto M, Miura I, Hirano M: **Clinical features and prognosis of de novo diffuse large B-cell lymphoma with t(14;18) and 8q24/c-MYC translocations**. *Leukemia* 2009, **23**(4):777-783.
48. Masque-Soler N, Szczepanowski M, Kohler CW, Spang R, Klapper W: **Molecular classification of mature aggressive B-cell lymphoma using digital multiplexed gene expression on formalin-fixed paraffin-embedded biopsy specimens**. *Blood* 2013, **122**(11):1985-1986.
49. Carey CD, Gusenleitner D, Chapuy B, Kovach AE, Kluk MJ, Sun HH, Crossland RE, Bacon CM, Rand V, Dal Cin P *et al*: **Molecular Classification of MYC-Driven B-Cell Lymphomas by Targeted Gene Expression Profiling of Fixed Biopsy Specimens**. *Journal of Molecular Diagnostics* 2015, **17**(1):19-30.
50. Richter J, Schlesner M, Hoffmann S, Kreuz M, Leich E, Burkhardt B, Rosolowski M, Ammerpohl O, Wagener R, Bernhart SH *et al*: **Recurrent mutation of the ID3 gene in Burkitt lymphoma identified by integrated genome, exome and transcriptome sequencing**. *Nat Genet* 2012, **44**(12):1316-1320.
51. Salaverria I, Martin-Guerrero I, Wagener R, Kreuz M, Kohler CW, Richter J, Pienkowska-Grela B, Adam P, Burkhardt B, Claviez A *et al*: **A recurrent 11q aberration pattern characterizes a subset of MYC-negative high-grade B-cell lymphomas resembling Burkitt lymphoma**. *Blood* 2014, **123**(8):1187-1198.

52. Schrader A, Bentink S, Spang R, Lenze D, Hummel M, Kuo M, Arrand JR, Murray PG, Trumper L, Kube D *et al*: **High Myc activity is an independent negative prognostic factor for diffuse large B cell lymphomas**. *Int J Cancer* 2012, **131**(4):E348-361.
53. Valera A, Lopez-Guillermo A, Cardesa-Salzmann T, Climent F, Gonzalez-Barca E, Mercadal S, Espinosa I, Novelli S, Briones J, Mate JL *et al*: **MYC protein expression and genetic alterations have prognostic impact in patients with diffuse large B-cell lymphoma treated with immunochemotherapy**. *Haematologica* 2013, **98**(10):1554-1562.
54. Zhou KG, Xu DM, Cao Y, Wang J, Yang YF, Huang M: **C-MYC Aberrations as Prognostic Factors in Diffuse Large B-cell Lymphoma: A Meta-Analysis of Epidemiological Studies**. *PLoS One* 2014, **9**(4).
55. Kojima M, Nishikii H, Takizawa J, Aoki S, Noguchi M, Chiba S, Ando K, Nakamura N: **MYC rearrangements are useful for predicting outcomes following rituximab and chemotherapy: multicenter analysis of Japanese patients with diffuse large B-cell lymphoma**. *Leukemia Lymphoma* 2013, **54**(10):2149-2154.
56. Puvvada S, Kendrick S, Rimsza L: **Molecular classification, pathway addiction, and therapeutic targeting in diffuse large B cell lymphoma**. *Cancer Genet-Ny* 2013, **206**(7-8):257-265.
57. Johnson NA, Slack GW, Savage KJ, Connors JM, Ben-Neriah S, Rogic S, Scott DW, Tan KL, Steidl C, Sehn LH *et al*: **Concurrent expression of MYC and BCL2 in diffuse large B-cell lymphoma treated with rituximab plus cyclophosphamide, doxorubicin, vincristine, and prednisone**. *J Clin Oncol* 2012, **30**(28):3452-3459.
58. Perry AM, Alvarado-Bernal Y, Laurini JA, Smith LM, Slack GW, Tan KL, Sehn LH, Fu K, Aoun P, Greiner TC *et al*: **MYC and BCL2 protein expression predicts survival in patients with diffuse large B-cell lymphoma treated with rituximab**. *Brit J Haematol* 2014, **165**(3):382-391.
59. Chang TC, Yu DN, Lee YS, Wentzel EA, Arking DE, West KM, Dang CV, Thomas-Tikhonenko A, Mendell JT: **Widespread microRNA repression by Myc contributes to tumorigenesis**. *Nature Genetics* 2008, **40**(1):43-50.
60. Calado DP, Sasaki Y, Godinho SA, Pellerin A, Kochert K, Sleckman BP, de Alboran IM, Janz M, Rodig S, Rajewsky K: **The cell-cycle regulator c-Myc is essential for the formation and maintenance of germinal centers**. *Nat Immunol* 2012, **13**(11):1092-1100.
61. Klein U, Dalla-Favera R: **Germinal centres: role in B-cell physiology and malignancy**. *Nat Rev Immunol* 2008, **8**(1):22-33.
62. Spender LC, Inman GJ: **Developments in Burkitt's lymphoma: novel cooperations in oncogenic MYC signaling**. *Cancer Manag Res* 2014, **6**:27-38.
63. McKeown MR, Bradner JE: **Therapeutic Strategies to Inhibit MYC**. *Csh Perspect Med* 2014, **4**(10).
64. Berg T: **Small-Molecule Modulators of c-Myc/Max and Max/Max Interactions**. *Curr Top Microbiol* 2011, **348**:139-149.

65. Albihn A, Johnsen JI, Henriksson MA: **MYC in Oncogenesis and as a Target for Cancer Therapies**. *Advances in Cancer Research, Vol 107* 2010, **107**:163-224.
66. Mertz JA, Conery AR, Bryant BM, Sandy P, Balasubramanian S, Mele DA, Bergeron L, Sims RJ: **Targeting MYC dependence in cancer by inhibiting BET bromodomains**. *P Natl Acad Sci USA* 2011, **108**(40):16669-16674.
67. Chen BJ, Wu YL, Tanaka Y, Zhang W: **Small Molecules Targeting c-Myc Oncogene: Promising Anti-Cancer Therapeutics**. *Int J Biol Sci* 2014, **10**(10):1084-1096.
68. Oostlander AE, Meijer GA, Ylstra B: **Microarray-based comparative genomic hybridization and its applications in human genetics**. *Clin Genet* 2004, **66**(6):488-495.
69. Pinkel D, Albertson DG: **Array comparative genomic hybridization and its applications in cancer**. *Nat Genet* 2005, **37 Suppl**:S11-17.
70. Craddock KJ, Lam WL, Tsao MS: **Applications of array-CGH for lung cancer**. *Methods Mol Biol* 2013, **973**:297-324.
71. Laskowska J, Szczepanek J, Styczynski J, Tretyn A: **Array comparative genomic hybridization in pediatric acute leukemias**. *Pediatr Hematol Oncol* 2013, **30**(8):677-687.
72. Sadikovic B, Park PC, Selvarajah S, Zielenska M: **Array comparative genomic hybridization in osteosarcoma**. *Methods Mol Biol* 2013, **973**:227-247.
73. Kanamori M, Sano A, Yasuda T, Hori T, Suzuki K: **Array-based comparative genomic hybridization for genomic-wide screening of DNA copy number alterations in aggressive bone tumors**. *J Exp Clin Cancer Res* 2012, **31**:100.
74. Sachidanandam R, Weissman D, Schmidt SC, Kakol JM, Stein LD, Marth G, Sherry S, Mullikin JC, Mortimore BJ, Willey DL *et al*: **A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms**. *Nature* 2001, **409**(6822):928-933.
75. Johnson AD, O'Donnell CJ: **An open access database of genome-wide association results**. *BMC Med Genet* 2009, **10**:6.
76. LaFramboise T: **Single nucleotide polymorphism arrays: a decade of biological, computational and technological advances**. *Nucleic Acids Res* 2009, **37**(13):4181-4193.
77. Macgregor PF: **Gene expression in cancer: the application of microarrays**. *Expert Rev Mol Diagn* 2003, **3**(2):185-200.
78. Michiels S, Koscielny S, Hill C: **Prediction of cancer outcome with microarrays: a multiple random validation strategy**. *Lancet* 2005, **365**(9458):488-492.
79. Marioni JC, Mason CE, Mane SM, Stephens M, Gilad Y: **RNA-Seq: an assessment of technical reproducibility and comparison with gene expression arrays**. *Genome Res* 2008, **18**(9):1509-1517.
80. Sirbu A, Kerr G, Crane M, Ruskin HJ: **RNA-Seq vs dual- and single-channel microarray data: sensitivity analysis for differential expression and clustering**. *PLoS One* 2012, **7**(12):e50986.
81. Finotello F, Di Camillo B: **Measuring differential gene expression with RNA-Seq: challenges and strategies for data analysis**. *Brief Funct Genomics* 2014.

82. Capobianco E: **RNA-Seq Data: A Complexity Journey**. *Comput Struct Biotechnol J* 2014, **11**(19):123-130.
83. Jones PA, Baylin SB: **The fundamental role of epigenetic events in cancer**. *Nat Rev Genet* 2002, **3**(6):415-428.
84. Beier V, Mund C, Hoheisel JD: **Monitoring methylation changes in cancer**. *Adv Biochem Eng Biotechnol* 2007, **104**:1-11.
85. Reinders J, Paszkowski J: **Bisulfite methylation profiling of large genomes**. *Epigenomics* 2010, **2**(2):209-220.
86. Schumacher A, Kapranov P, Kaminsky Z, Flanagan J, Assadzadeh A, Yau P, Virtanen C, Winegarden N, Cheng J, Gingeras T *et al*: **Microarray-based DNA methylation profiling: technology and applications**. *Nucleic Acids Res* 2006, **34**(2):528-542.
87. Park PJ: **ChIP-Seq: advantages and challenges of a maturing technology**. *Nat Rev Genet* 2009, **10**(10):669-680.
88. Cedar H, Bergman Y: **Linking DNA methylation and histone modification: patterns and paradigms**. *Nat Rev Genet* 2009, **10**(5):295-304.
89. Park YJ, Claus R, Weichenhan D, Plass C: **Genome-wide epigenetic modifications in cancer**. *Prog Drug Res* 2011, **67**:25-49.
90. Chaerkady R, Pandey A: **Quantitative proteomics for identification of cancer biomarkers**. *Proteomics Clin Appl* 2007, **1**(9):1080-1089.
91. Boja ES, Rodriguez H: **Mass spectrometry-based targeted quantitative proteomics: achieving sensitive and reproducible detection of proteins**. *Proteomics* 2012, **12**(8):1093-1110.
92. LiEBLer DC, Zimmerman LJ: **Targeted quantitation of proteins by mass spectrometry**. *Biochemistry* 2013, **52**(22):3797-3806.
93. Lindon JC, Nicholson JK: **Spectroscopic and statistical techniques for information recovery in metabonomics and metabolomics**. *Annu Rev Anal Chem (Palo Alto Calif)* 2008, **1**:45-69.
94. Somashekar BS, Kamarajan P, Danciu T, Kapila YL, Chinnaiyan AM, Rajendiran TM, Ramamoorthy A: **Magic angle spinning NMR-based metabolic profiling of head and neck squamous cell carcinoma tissues**. *J Proteome Res* 2011, **10**(11):5232-5241.
95. Poullet JB, Martinez-Bisbal MC, Valverde D, Monleon D, Celda B, Arus C, Van Huffel S: **Quantification and classification of high-resolution magic angle spinning data for brain tumor diagnosis**. *Conf Proc IEEE Eng Med Biol Soc* 2007, **2007**:5407-5410.
96. Hayes DF: **OMICS-based personalized oncology: if it is worth doing, it is worth doing well!** *BMC Med* 2013, **11**:221.
97. Ransohoff DF: **The process to discover and develop biomarkers for cancer: a work in progress**. *J Natl Cancer Inst* 2008, **100**(20):1419-1420.
98. McShane LM, Hayes DF: **Publication of tumor marker research results: the necessity for complete and transparent reporting**. *J Clin Oncol* 2012, **30**(34):4223-4232.
99. Arizmendi C, Sierra DA, Vellido A, Romero E: **Brain tumour classification using Gaussian decomposition and neural networks**. *Conf Proc IEEE Eng Med Biol Soc* 2011, **2011**:5645-5648.
100. McShane LM, Cavenagh MM, Lively TG, Eberhard DA, Bigbee WL, Williams PM, Mesirov JP, Polley MY, Kim KY, Tricoli JV *et al*: **Criteria**

for the use of omics-based predictors in clinical trials: explanation and elaboration. *BMC Med* 2013, **11**:220.
101. Rew DA: **DNA microarray technology in cancer research**. *Eur J Surg Oncol* 2001, **27**(5):504-508.
102. Frolov AE, Godwin AK, Favorova OO: **[Differential gene expression analysis by DNA microarrays technology and its application in molecular oncology]**. *Mol Biol (Mosk)* 2003, **37**(4):573-584.
103. Clarke PA, te Poele R, Wooster R, Workman P: **Gene expression microarray analysis in cancer biology, pharmacology, and drug development: progress and potential**. *Biochem Pharmacol* 2001, **62**(10):1311-1336.
104. Afshari CA, Nuwaysir EF, Barrett JC: **Application of complementary DNA microarray technology to carcinogen identification, toxicology, and drug safety evaluation**. *Cancer Res* 1999, **59**(19):4759-4760.
105. Xu Y, Duanmu H, Chang Z, Zhang S, Li Z, Liu Y, Li K, Qiu F, Li X: **The application of gene co-expression network reconstruction based on CNVs and gene expression microarray data in breast cancer**. *Mol Biol Rep* 2012, **39**(2):1627-1637.
106. Bouma G, Baggen JM, van Bodegraven AA, Mulder CJ, Kraal G, Zwiers A, Horrevoets AJ, van der Pouw Kraan CT: **Thiopurine treatment in patients with Crohn's disease leads to a selective reduction of an effector cytotoxic gene expression signature revealed by whole-genome expression profiling**. *Mol Immunol* 2013, **54**(3-4):472-481.
107. Yoshida S, Ishibashi T: **Development of novel molecular targeting therapy for diabetic retinopathy based on genome-wide gene expression profiling**. *Fukuoka Igaku Zasshi* 2013, **104**(8):240-247.
108. O'Brien MA, Costin BN, Miles MF: **Using Genome-Wide Expression Profiling to Define Gene Networks Relevant to the Study of Complex Traits: From RNA Integrity to Network Topology**. *Int Rev Neurobiol* 2012, **104**:91-133.
109. Macgregor PF, Squire JA: **Application of microarrays to the analysis of gene expression in cancer**. *Clin Chem* 2002, **48**(8):1170-1177.
110. Schena M, Shalon D, Davis RW, Brown PO: **Quantitative monitoring of gene expression patterns with a complementary DNA microarray**. *Science* 1995, **270**(5235):467-470.
111. Lockhart DJ, Dong H, Byrne MC, Follettie MT, Gallo MV, Chee MS, Mittmann M, Wang C, Kobayashi M, Horton H *et al*: **Expression monitoring by hybridization to high-density oligonucleotide arrays**. *Nat Biotechnol* 1996, **14**(13):1675-1680.
112. Frantz S: **An array of problems**. *Nat Rev Drug Discov* 2005, **4**(5):362-363.
113. Irizarry RA, Warren D, Spencer F, Kim IF, Biswal S, Frank BC, Gabrielson E, Garcia JG, Geoghegan J, Germino G *et al*: **Multiple-laboratory comparison of microarray platforms**. *Nat Methods* 2005, **2**(5):345-350.
114. Shi L, Reid LH, Jones WD, Shippy R, Warrington JA, Baker SC, Collins PJ, de Longueville F, Kawasaki ES, Lee KY *et al*: **The MicroArray Quality Control (MAQC) project shows inter- and**

intraplatform reproducibility of gene expression measurements. *Nat Biotechnol* 2006, **24**(9):1151-1161.

115. Chen JJ, Hsueh HM, Delongchamp RR, Lin CJ, Tsai CA: **Reproducibility of microarray data: a further analysis of microarray quality control (MAQC) data**. *BMC Bioinformatics* 2007, **8**:412.
116. Williams PM, Li R, Johnson NA, Wright G, Heath JD, Gascoyne RD: **A novel method of amplification of FFPET-derived RNA enables accurate disease classification with microarrays**. *J Mol Diagn* 2010, **12**(5):680-686.
117. Wang J: **Computational biology of genome expression and regulation--a review of microarray bioinformatics**. *J Environ Pathol Toxicol Oncol* 2008, **27**(3):157-179.
118. Quackenbush J: **Computational approaches to analysis of DNA microarray data**. *Yearb Med Inform* 2006:91-103.
119. Kerns RT, Miles MF: **Microarray analysis of ethanol-induced changes in gene expression**. *Methods Mol Biol* 2008, **447**:395-410.
120. Smyth GK, Speed T: **Normalization of cDNA microarray data**. *Methods* 2003, **31**(4):265-273.
121. Hijazi H, Chan C: **A classification framework applied to cancer gene expression profiles**. *J Healthc Eng* 2013, **4**(2):255-283.
122. Selvaraj S, Natarajan J: **Microarray data analysis and mining tools**. *Bioinformation* 2011, **6**(3):95-99.
123. Raspe E, Decraene C, Berx G: **Gene expression profiling to dissect the complexity of cancer biology: pitfalls and promise**. *Semin Cancer Biol* 2012, **22**(3):250-260.
124. Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B: **Mapping and quantifying mammalian transcriptomes by RNA-Seq**. *Nat Methods* 2008, **5**(7):621-628.
125. Hurd PJ, Nelson CJ: **Advantages of next-generation sequencing versus the microarray in epigenetic research**. *Brief Funct Genomic Proteomic* 2009, **8**(3):174-183.
126. McGowan P, Nelles N, Wimmer J, Williams D, Wen JG, Li M, Ewton A, Curry C, Zu YL, Sheehan A *et al*: **Differentiating Between Burkitt Lymphoma and CD10+Diffuse Large B-Cell Lymphoma The Role of Commonly Used Flow Cytometry Cell Markers and the Application of a Multiparameter Scoring System**. *American Journal of Clinical Pathology* 2012, **137**(4):665-670.
127. Yeh RG, Lin CW, Abbod MF, Shieh JS: **Two-dimensional matrix algorithm using detrended fluctuation analysis to distinguish Burkitt and diffuse large B-cell lymphoma**. *Comput Math Methods Med* 2012, **2012**:947191.
128. Dave SS, Fu K, Wright GW, Lam LT, Kluin P, Boerma EJ, Greiner TC, Weisenburger DD, Rosenwald A, Ott G *et al*: **Molecular diagnosis of Burkitt's lymphoma**. *N Engl J Med* 2006, **354**(23):2431-2442.
129. Hummel M, Bentink S, Berger H, Klapper W, Wessendorf S, Barth TF, Bernd HW, Cogliatti SB, Dierlamm J, Feller AC *et al*: **A biologic definition of Burkitt's lymphoma from transcriptional and genomic profiling**. *N Engl J Med* 2006, **354**(23):2419-2430.

130. Martin-Subero JI, Odero MD, Hernandez R, Cigudosa JC, Agirre X, Saez B, Sanz-Garcia E, Ardanaz MT, Novo FJ, Gascoyne RD *et al*: **Amplification of IGH/MYC fusion in clinically aggressive IGH/BCL2-positive germinal center B-cell lymphomas**. *Genes Chromosomes Cancer* 2005, **43**(4):414-423.
131. Barrans S, Crouch S, Smith A, Turner K, Owen R, Patmore R, Roman E, Jack A: **Rearrangement of MYC is associated with poor prognosis in patients with diffuse large B-cell lymphoma treated in the era of rituximab**. *J Clin Oncol* 2010, **28**(20):3360-3365.
132. Horn H, Ziepert M, Becher C, Barth TF, Bernd HW, Feller AC, Klapper W, Hummel M, Stein H, Hansmann ML *et al*: **MYC status in concert with BCL2 and BCL6 expression predicts outcome in diffuse large B-cell lymphoma**. *Blood* 2013, **121**(12):2253-2263.
133. Gearhart J, Pashos EE, Prasad MK: **Pluripotency redux - Advances in stem-cell research**. *New Engl J Med* 2007, **357**(15):1469-1472.
134. Chambers JM: **Software for Data Analysis: Programming with R**. *Stat Comput Ser* 2008:1-498.
135. Gentleman RC, Carey VJ, Bates DM, Bolstad B, Dettling M, Dudoit S, Ellis B, Gautier L, Ge YC, Gentry J *et al*: **Bioconductor: open software development for computational biology and bioinformatics**. *Genome Biol* 2004, **5**(10).
136. Edgar R, Domrachev M, Lash AE: **Gene Expression Omnibus: NCBI gene expression and hybridization array data repository**. *Nucleic Acids Research* 2002, **30**(1):207-210.
137. Barrett T, Wilhite SE, Ledoux P, Evangelista C, Kim IF, Tomashevsky M, Marshall KA, Phillippy KH, Sherman PM, Holko M *et al*: **NCBI GEO: archive for functional genomics data sets-update**. *Nucleic Acids Research* 2013, **41**(D1):D991-D995.
138. Quackenbush J: **Microarray data normalization and transformation**. *Nat Genet* 2002, **32 Suppl**:496-501.
139. Bolstad BM, Irizarry RA, Astrand M, Speed TP: **A comparison of normalization methods for high density oligonucleotide array data based on variance and bias**. *Bioinformatics* 2003, **19**(2):185-193.
140. Huber W, von Heydebreck A, Sultmann H, Poustka A, Vingron M: **Variance stabilization applied to microarray data calibration and to the quantification of differential expression**. *Bioinformatics* 2002, **18 Suppl 1**:S96-104.
141. Ambroise J, Bearzatto B, Robert A, Govaerts B, Macq B, Gala JL: **Impact of the spotted microarray preprocessing method on fold-change compression and variance stability**. *BMC Bioinformatics* 2011, **12**:413.
142. Hautaniemi S, Lehmussola A, Yli-Harja O: **DNA microarray data preprocessing**. *Isccsp : 2004 First International Symposium on Control, Communications and Signal Processing* 2004:751-754.
143. Gautier L, Cope L, Bolstad BM, Irizarry RA: **affy--analysis of Affymetrix GeneChip data at the probe level**. *Bioinformatics* 2004, **20**(3):307-315.
144. Du P, Kibbe WA, Lin SM: **lumi: a pipeline for processing Illumina microarray**. *Bioinformatics* 2008, **24**(13):1547-1548.

145. Shabalin AA, Tjelmeland H, Fan C, Perou CM, Nobel AB: **Merging two gene-expression studies via cross-platform normalization**. *Bioinformatics* 2008, **24**(9):1154-1160.
146. Benito M, Parker J, Du Q, Wu J, Xiang D, Perou CM, Marron JS: **Adjustment of systematic microarray data biases**. *Bioinformatics* 2004, **20**(1):105-114.
147. Rudy J, Valafar F: **Empirical comparison of cross-platform normalization methods for gene expression data**. *BMC Bioinformatics* 2011, **12**:467.
148. Eisen MB, Spellman PT, Brown PO, Botstein D: **Cluster analysis and display of genome-wide expression patterns**. *Proc Natl Acad Sci U S A* 1998, **95**(25):14863-14868.
149. Soukas A, Cohen P, Socci ND, Friedman JM: **Leptin-specific patterns of gene expression in white adipose tissue**. *Genes Dev* 2000, **14**(8):963-980.
150. Liu Q, Sung AH, Chen Z, Liu J, Chen L, Qiao M, Wang Z, Huang X, Deng Y: **Gene selection and classification for cancer microarray data based on machine learning and similarity measures**. *BMC Genomics* 2011, **12 Suppl 5**:S1.
151. Dessi N, Pascariello E, Pes B: **A comparative analysis of biomarker selection techniques**. *Biomed Res Int* 2013, **2013**:387673.
152. Inza I, Sierra B, Blanco R: **Gene selection by sequential search wrapper approaches in microarray cancer class prediction**. *J Intell Fuzzy Syst* 2002, **12**(1):25-33.
153. Tusher VG, Tibshirani R, Chu G: **Significance analysis of microarrays applied to the ionizing radiation response (vol 98, pg 5116, 2001)**. *P Natl Acad Sci USA* 2001, **98**(18):10515-10515.
154. Li GCJ: **"Significance Analysis of Microarrays" Users guide and technical document**.
155. Ressom HW, Varghese RS, Zhang Z, Xuan J, Clarke R: **Classification algorithms for phenotype prediction in genomics and proteomics**. *Front Biosci* 2008, **13**:691-708.
156. Satagopan JM, Panageas KS: **A statistical perspective on gene expression data analysis**. *Stat Med* 2003, **22**(3):481-499.
157. Tibshirani R, Hastie T, Narasimhan B, Chu G: **Diagnosis of multiple cancer types by shrunken centroids of gene expression**. *P Natl Acad Sci USA* 2002, **99**(10):6567-6572.
158. Rumelhart DE, Hinton GE, Williams RJ: **Learning Representations by Back-Propagating Errors**. *Nature* 1986, **323**(6088):533-536.
159. Shawe-Taylor NCJ: **An introduction to support vector machines and other kernel-based learning methods**. 2000.
160. Chih-Wei Hsu C-CC, Chih-Jen Lin: **A Practical Guide to Support Vector Classification**. 2003.
161. Corinna Cortes VV: **Support-vector networks**. *Machine Learning* 1995, **20**(3):273-297.
162. Flynn R: **Survival analysis**. *J Clin Nurs* 2012, **21**(19-20):2789-2797.
163. Bradburn MJ, Clark TG, Love SB, Altman DG: **Survival analysis part II: multivariate data analysis--an introduction to concepts and methods**. *Br J Cancer* 2003, **89**(3):431-436.

164. Kaplan EL, Meier P: **Nonparametric-Estimation from Incomplete Observations**. *J Am Stat Assoc* 1958, **53**(282):457-481.
165. Peto R, Pike MC, Armitage P, Breslow NE, Cox DR, Howard SV, Mantel N, Mcpherson K, Peto J, Smith PG: **Design and Analysis of Randomized Clinical-Trials Requiring Prolonged Observation of Each Patient .2. Analysis and Examples**. *Brit J Cancer* 1977, **35**(1):1-39.
166. Campbell F, Smith RA, Whelan P, Sutton R, Raraty M, Neoptolemos JP, Ghaneh P: **Classification of R1 resections for pancreatic cancer: the prognostic relevance of tumour involvement within 1 mm of a resection margin**. *Histopathology* 2009, **55**(3):277-283.
167. Ashburner M, Ball CA, Blake JA, Butler H, Cherry JM, Corradi J, Dolinski K, Eppig JT, Harris M, Hill DP *et al*: **Creating the gene ontology resource: Design and implementation**. *Genome Research* 2001, **11**(8):1425-1433.
168. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT *et al*: **Gene Ontology: tool for the unification of biology**. *Nature Genetics* 2000, **25**(1):25-29.
169. Yip YL, Scheib H, Diemand AV, Gattiker A, Famiglietti LM, Gasteiger E, Bairoch A: **The Swiss-Prot variant page and the ModSNP database: A resource for sequence and structure information on human protein variants**. *Hum Mutat* 2004, **23**(5):464-470.
170. Bairoch A, Apweiler R: **The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000**. *Nucleic Acids Research* 2000, **28**(1):45-48.
171. Benson DA, Clark K, Karsch-Mizrachi I, Lipman DJ, Ostell J, Sayers EW: **GenBank**. *Nucleic Acids Research* 2014, **42**(D1):D32-D37.
172. Barker WC, Garavelli JS, Huang HZ, McGarvey PB, Orcutt BC, Srinivasarao GY, Xiao CL, Yeh LSL, Ledley RS, Janda JF *et al*: **The Protein Information Resource (PIR)**. *Nucleic Acids Research* 2000, **28**(1):41-44.
173. Zeeberg BR, Qin HY, Narasimhan S, Sunshine M, Cao H, Kane DW, Reimers M, Stephens RM, Bryant D, Burt SK *et al*: **High-Throughput GoMiner, an 'industrial-strength' integrative gene ontology tool for interpretation of multiple-microarray experiments, with application to studies of Common Variable Immune Deficiency (CVID)**. *BMC Bioinformatics* 2005, **6**.
174. Beissbarth T, Speed TP: **GOstat: find statistically overrepresented Gene Ontologies within a group of genes**. *Bioinformatics* 2004, **20**(9):1464-1465.
175. Masseroli M, Galati O, Pinciroli F: **GFINDer: genetic disease and phenotype location statistical analysis and mining of dynamically annotated gene lists**. *Nucleic Acids Research* 2005, **33**:W717-W723.
176. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES *et al*: **Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles**. *Proc Natl Acad Sci U S A* 2005, **102**(43):15545-15550.

177. Dennis G, Sherman BT, Hosack DA, Yang J, Gao W, Lane HC, Lempicki RA: **DAVID: Database for annotation, visualization, and integrated discovery**. *Genome Biol* 2003, **4**(9).
178. Liberzon A: **A Description of the Molecular Signatures Database (MSigDB) Web Site**. *Stem Cell Transcriptional Networks: Methods and Protocols* 2014, **1150**:153-160.
179. Sherman BT, Huang DW, Tan QN, Guo YJ, Bour S, Liu D, Stephens R, Baseler MW, Lane HC, Lempicki RA: **DAVID Knowledgebase: a gene-centered database integrating heterogeneous gene annotation resources to facilitate high-throughput gene functional analysis**. *BMC Bioinformatics* 2007, **8**.
180. Chan WK, Mak HK, Huang B, Yeung DW, Kwong DL, Khong PL: **Nasopharyngeal carcinoma: relationship between 18F-FDG PET-CT maximum standardized uptake value, metabolic tumour volume and total lesion glycolysis and TNM classification**. *Nucl Med Commun* 2010, **31**(3):206-210.
181. Deffenbacher KE, Iqbal J, Liu Z, Fu K, Chan WC: **Recurrent chromosomal alterations in molecularly classified AIDS-related lymphomas: an integrated analysis of DNA copy number and gene expression**. *J Acquir Immune Defic Syndr* 2010, **54**(1):18-26.
182. Piccaluga PP, De Falco G, Kustagi M, Gazzola A, Agostinelli C, Tripodo C, Leucci E, Onnis A, Astolfi A, Sapienza MR *et al*: **Gene expression analysis uncovers similarity and differences among Burkitt lymphoma subtypes**. *Blood* 2011, **117**(13):3596-3608.
183. Klapper W, Szczepanowski M, Burkhardt B, Berger H, Rosolowski M, Bentink S, Schwaenen C, Wessendorf S, Spang R, Moller P *et al*: **Molecular profiling of pediatric mature B-cell lymphoma treated in population-based prospective clinical trials**. *Blood* 2008, **112**(4):1374-1381.
184. Carlson M: **hgu133a.db: Affymetrix Human Genome U133 Set annotation data (chip hgu133a)**.
185. Gray KA, Daugherty LC, Gordon SM, Seal RL, Wright MW, Bruford EA: **Genenames.org: the HGNC resources in 2013**. *Nucleic Acids Research* 2013, **41**(D1):D545-D552.
186. Geiss GK, Bumgarner RE, Birditt B, Dahl T, Dowidar N, Dunaway DL, Fell HP, Ferree S, George RD, Grogan T *et al*: **Direct multiplexed measurement of gene expression with color-coded probe pairs**. *Nat Biotechnol* 2008, **26**(3):317-325.
187. **Whole-Genome DASL HT Assay for Expresseion Profiling in FFPE Samples** [http://support.illumina.com/content/dam/illumina-marketing/documents/products/datasheets/datasheet_whole_genome_dasl_ht.pdf]
188. April C, Klotzle B, Royce T, Wickham-Garcia E, Boyaniwsky T, Izzo J, Cox D, Jones W, Rubio R, Holton K *et al*: **Whole-Genome Gene Expression Profiling of Formalin-Fixed, Paraffin-Embedded Tissue Samples**. *PLoS One* 2009, **4**(12).
189. Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, Smyth GK: **limma powers differential expression analyses for RNA-Sequencing and microarray studies**. *Nucleic Acids Res* 2015.
190. Leisch F: **Creating R Packages: A Tutorial**. 2008.

191. Green TM, Nielsen O, de Stricker K, Xu-Monette ZY, Young KH, Moller MB: **High levels of nuclear MYC protein predict the presence of MYC rearrangement in diffuse large B-cell lymphoma**. *Am J Surg Pathol* 2012, **36**(4):612-619.
192. Salaverria I, Philipp C, Oschlies I, Kohler CW, Kreuz M, Szczepanowski M, Burkhardt B, Trautmann H, Gesk S, Andrusiewicz M *et al*: **Translocations activating IRF4 identify a subtype of germinal center-derived B-cell lymphoma affecting predominantly children and young adults**. *Blood* 2011, **118**(1):139-147.
193. Cardesa-Salzmann TM, Colomo L, Gutierrez G, Chan WC, Weisenburger D, Climent F, Gonzalez-Barca E, Mercadal S, Arenillas L, Serrano S *et al*: **High microvessel density determines a poor outcome in patients with diffuse large B-cell lymphoma treated with rituximab plus chemotherapy**. *Haematologica* 2011, **96**(7):996-1001.
194. Frei E, Visco C, Xu-Monette ZY, Dirnhofer S, Dybkaer K, Orazi A, Bhagat G, Hsi ED, van Krieken JH, Ponzoni M *et al*: **Addition of rituximab to chemotherapy overcomes the negative prognostic impact of cyclin E expression in diffuse large B-cell lymphoma**. *J Clin Pathol* 2013, **66**(11):956-961.
195. Hu S, Xu-Monette ZY, Tzankov A, Green T, Wu L, Balasubramanyam A, Liu WM, Visco C, Li Y, Miranda RN *et al*: **MYC/BCL2 protein coexpression contributes to the inferior survival of activated B-cell subtype of diffuse large B-cell lymphoma and demonstrates high-risk gene expression signatures: a report from The International DLBCL Rituximab-CHOP Consortium Program**. *Blood* 2013, **121**(20):4021-4031; quiz 4250.
196. Compagno M, Lim WK, Grunn A, Nandula SV, Brahmachary M, Shen Q, Bertoni F, Ponzoni M, Scandurra M, Califano A *et al*: **Mutations of multiple genes cause deregulation of NF-kappaB in diffuse large B-cell lymphoma**. *Nature* 2009, **459**(7247):717-721.
197. Monti S, Chapuy B, Takeyama K, Rodig SJ, Hao Y, Yeda KT, Inguilizian H, Mermel C, Currie T, Dogan A *et al*: **Integrative analysis reveals an outcome-associated and targetable pattern of p53 and cell cycle deregulation in diffuse large B cell lymphoma**. *Cancer Cell* 2012, **22**(3):359-372.
198. Monti S, Savage KJ, Kutok JL, Feuerhake F, Kurtin P, Mihm M, Wu BY, Pasqualucci L, Neuberg D, Aguiar RCT *et al*: **Molecular profiling of diffuse large B-cell lymphoma identifies robust subtypes including one characterized by host inflammatory response**. *Blood* 2005, **105**(5):1851-1861.
199. Cananzi FC, Judson I, Lorenzi B, Benson C, Mudan S: **Multidisciplinary care of gastrointestinal stromal tumour: a review and a proposal for a pre-treatment classification**. *Eur J Surg Oncol* 2013, **39**(11):1171-1178.
200. Pal S, Yun R, Datta A, Lacomis L, Erdjument-Bromage H, Kumar J, Tempst P, Sif S: **mSin3A/histone deacetylase 2- and PRMT5-containing Brg1 complex is involved in transcriptional repression of the Myc target gene cad**. *Mol Cell Biol* 2003, **23**(21):7475-7487.

201. Tanaka H, Hoshikawa Y, Oh-hara T, Koike S, Naito M, Noda T, Arai H, Tsuruo T, Fujita N: **PRMT5, a novel TRAIL receptor-binding protein, inhibits TRAIL-induced apoptosis via nuclear factor-kappaB activation**. *Mol Cancer Res* 2009, **7**(4):557-569.
202. Kanno Y, Inajima J, Kato S, Matsumoto M, Tokumoto C, Kure Y, Inouye Y: **Protein arginine methyltransferase 5 (PRMT5) is a novel coactivator of constitutive androstane receptor (CAR)**. *Biochem Bioph Res Co* 2015, **459**(1):143-147.
203. Yang F, Wang J, Ren HY, Jin J, Wang AL, Sun LL, Diao KX, Wang EH, Mi XY: **Proliferative role of TRAF4 in breast cancer by upregulating PRMT5 nuclear expression**. *Tumour Biol* 2015.
204. Stopa N, Krebs JE, Shechter D: **The PRMT5 arginine methyltransferase: many roles in development, cancer and beyond**. *Cell Mol Life Sci* 2015, **72**(11):2041-2059.
205. Chung JH, Karkhanis V, Tae S, Yan FT, Smith P, Ayers LW, Agostinelli C, Pileri S, Denis GV, Baiocchi RA *et al*: **Protein Arginine Methyltransferase 5 (PRMT5) Inhibition Induces Lymphoma Cell Death through Reactivation of the Retinoblastoma Tumor Suppressor Pathway and Polycomb Repressor Complex 2 (PRC2) Silencing**. *J Biol Chem* 2013, **288**(49):35534-35547.
206. Li Y, Chitnis N, Nakagawa H, Kita Y, Natsugoe S, Yang Y, Li ZH, Wasik M, Klein-Szanto AJP, Rustgi AK *et al*: **PRMT5 Is Required for Lymphomagenesis Triggered by Multiple Oncogenic Drivers**. *Cancer Discov* 2015, **5**(3):288-303.
207. Wang L, Pal S, Sif S: **Protein arginine methyltransferase 5 suppresses the transcription of the RB family of tumor suppressors in leukemia and lymphoma cells**. *Mol Cell Biol* 2008, **28**(20):6262-6277.
208. Yan FT, Alinari L, Lustberg ME, Martin LK, Cordero-Nieves HM, Banasavadi-Siddegowda Y, Virk S, Barnholtz-Sloan J, Bell EH, Wojton J *et al*: **Genetic Validation of the Protein Arginine Methyltransferase PRMT5 as a Candidate Therapeutic Target in Glioblastoma**. *Cancer Research* 2014, **74**(6):1752-1765.
209. Alinari L, Mahasenan KV, Yan F, Karkhanis V, Chung JH, Smith EM, Quinion C, Smith PL, Kim L, Patton JT *et al*: **Selective inhibition of protein arginine methyltransferase 5 blocks initiation and maintenance of B-cell transformation**. *Blood* 2015.

Appendix A

The genes in each tested gene lists

Gene10:

SMARCA4, TCF3, CTSH, STAT3, BCL2A1, CD44, NASP, RNASEH2B, PRKAR2B, PRDM10

Gene 21:

SMARCA4, SLC35E3, SSBP2, MME, RGCC, BMP7, BACH2, RFC3, DLEU1, TERT, TCF3, ID3, MDFIC, BCL2A1, NFKBIA, FNBP1, CTSH, CD40, STAT3, CD44, CFLAR

Gene 28:

MARCA4, SLC35E3, SSBP2, MME, RGCC, BMP7, BACH2, RFC3, DLEU1, TERT, TCF3, ID3, TCL6, LEF1, C7orf10, SOX11, TUBA1A, MDFIC, S100A11, BCL2A1, NFKBIA, FNBP1, CTSH, CD40, STAT3, CD44, CFLAR, BCL3

Gene 60:

SMARCA4, SLC35E3, RFC3, DLEU1, TERT, NFKBIA, CD40, STAT3, BCL2A1, AHR, FNBP1, EBI3, TCF3, CFLAR, CTSH, CD44, MDFIC, SAMSN1, NREP, ID3, IL21R, BATF, TNFRSF1B, CCR7, BMP7, LY75, FAS, RBFOX2, LMO2, ENTPD1, RGCC, HCK, BCL3, SELL, RAB7L1, SSBP2, CXCL13, PBK, SOX11, BACH2, PTGDS, NCF2, TCL1A, VPREB3, BCL2, LEF1, CCND2, MME, CCL19, ALOX5, CXCL9, CD24, GZMB, CLU, CYB5R2, IGJ, AUTS2, MYBL1, TNFRSF17, BANK1

Gene 173:

SMARCA4, NDC1, ARHGEF18, GLRX5, SAC3D1, CDC25B, TAPT1, MAZ, RFC3, NREP, RGCC, BYSL, HNRNPAB, VRK1, DHFR, PNN, ID3, WHSC1, RBFOX2, TTLL12, HNRNPU, RANBP1, CDC7, GJC1, DEPDC1, ZWILCH, POLE3, TCL1A, CDKN2C, GTF3A, KIAA0226L, SLC35E3, SSBP2, NT5DC2, RFC4, GMNN, ARHGAP19, PRR11, AUTS2, PPP1R14B, HIP1R, GNA13, MYBL1, PTTG1, BMP7, DPY19L2P2, TMEM97, CKS1B, ITPR3, MSH6, NUDT21, NUP205, MPHOSPH9, CDK13, CBX2, DUT, RALBP1,

BACH2, OXCT1, TTK, MIF, CDC6, MME, LTBP1, CDK4, UBE2S, CCNB2, ALOX5, ANAPC15, EIF2AK3, VPREB3, MAP3K4, TOP1, MRPS2, TCF3, CD320, NUDT6, YTHDF1, SNRPA1, JAK2, CMTM6, ZSWIM8, IL2RG, NUP62, CCND1, CD83, SBNO2, CXCR3, SGK1, TLR1, HLA-F, CD58, TMBIM6, IL16, ST3GAL5, IL18, PIM1, LIMD1, TLR2, TP63, DTX4, ADTRP, CASP8, IL10RA, DENND3, ITGB2, CD40, HLA-E, CYB5R2, RIN2, DOCK10, AHNAK, CLIP2, TNFRSF1B, PTPN1, RASGRF1, PIM2, SNX11, SLA, TPP1, NECAP2, TRAF1, PTGIR, RAB7L1, ENTPD1, HCK, ATXN1, TNFSF10, HLA-G, IL6R, CTSH, LCP2, ATP6V0E1, ARPC1B, CCR7, KYNU, SLAMF1, MDFIC, FAS, JAK3, BATF, LMO2, ICAM1, LY75, NFKBIA, BATF3, TNFAIP3, FNBP1, CFLAR, MAP3K8, AHR, CD44, SAMSN1, EBI3, STAT3, BCL2A1, TERT, DLEU1, MYC, HDGF, TRIP13, BUB1B, GRSF1, LRPPRC, SFPQ, SRPK1, HDAC2, HMGB1, MAPKAPK5, CSTF3, GLO1, NOLC1.

Appendix B

Details of 48 clinically diagnosed DLBCL cases with *MYC* translocation

Table: Detailed clinical information of 48 *MYC*-rearranged DLBCL cases

Sample.ID [1]	BL.prob[2]	Treatment[3]	Survival/ Response[4]	BCL2, BCL6 rearrangement
14 cases with MYC-rearrangement classified as BL by the GEP classifier				
H12726/06	0.52	R-CHOP	Alive	BCL2 rearranged
H2123/11	0.617	R-CHOP	Alive	BCL2, BCL6 rearranged
RCH_H14473/10	0.75	R-CHOP	Complete remission	BCL2, BCL6 rearranged
RCH_21356	0.665	R-CHOP	Persistent response	BCL2 rearranged
H10150/04	0.716	R-CHOP	Died	BCL2 rearranged
H1032/07	0.662	R-CHOP	Died	BCL2, BCL6 rearranged
H11330/06	0.745	R-CHOP	Died	BCL2 rearranged

H23841/11	0.738	R-CHOP	Died	BCL2 rearranged
H2470/05	0.561	R-CHOP	Died	BCL2 rearranged
H348/05	0.552	R-CHOP	Died	BCL2 rearranged
LY10_1069	0.777	CODOX-M/IVAC	Died	BCL2 rearranged, BCL6 normal
H11977/04	0.674	unknown	unknown	BCL2 rearranged
H19874_13	0.734	unknown	unknown	BCL2 rearranged
H3218/05	0.711	unknown	unknown	BCL2, BCL6 rearranged

34 cases with MYC-rearrangement classified as DLBCL by the GEP classifier

H5694/12	0.029	R-CHOP	Alive	BCL6 rearranged
H10453/04	0.017	R-CHOP	Alive	BCL6 rearranged
H11800/04	0.07	R-CHOP	Alive	BCL6 rearranged
H682/12	0.126	R-CHOP	Alive	BCL2 amplified
H24187/11	0.168	R-CHOP	Alive	BCL2 rearranged

H2715/12	0.008	R-CHOP	Alive	BCL2 rearranged
H6062/06	0.032	R-CHOP	Alive	BCL2 rearranged
H7755/12	0.136	R-CHOP	Alive	BCL2 rearranged
H9553/04	0.01	R-CHOP	Alive	BCL2 rearranged
H4658/04	0.023	R-CHOP	Alive	BCL2, BCL6 normal
LY10_1130	0.016	R-CHOP	Alive	BCL2 normal, BCL6 rearranged
LY10_1144	0.037	R-CHOP	Alive	BCL2 normal, BCL6 rearranged
RCH_H19096/10	0.086	R-CHOP	Complete remission	BCL6 rearranged
RCH_H19093/10	0.148	R-CHOP	Complete remission	BCL2 rearranged
RCH_H14275/10	0.19	R-CHOP	Complete remission	BCL2 rearranged
RCH_20462	0.162	R-CHOP	Complete remission	BCL2, BCL6 normal
RCH_H13616/10	0.375	R-CHOP	Complete remission	BCL2, BCL6 normal
RCH_H1144_11	0.008	R-CHOP	Persistent disease (die)	BCL6 rearranged

LY10_1135	0.071	R-CHOP	Died	BCL2 rearranged, BCL6 normal
LY10_1065	0.15	R-CHOP	Died	BCL2, BCL6 normal
H12035/07	0.245	R-CHOP	Died	BCL2 rearranged
H1102/06	0.009	R-CHOP	Died	BCL2, BCL6 normal
H12162/04	0.195	R-CHOP	Died	BCL2, BCL6 normal
H3321/11	0.088	R-CHOP	Died	BCL2, BCL6 normal
H3902/06	0.02	R-CHOP	Died	BCL2, BCL6 normal
LY10_1077	0.5	CODOX-M/IVAC	Died	BCL2 rearranged, BCL6 normal
H3251/06	0.034	No active Treatment	Died	BCL2 rearranged
H1256/05	0.015	No active Treatment	Died	BCL2 rearranged
H10073/04	0.082	No active Treatment	Died	BCL6 rearranged, BCL2 amplified
H17622/10	0.015	No active Treatment	Died	BCL2, BCL6 normal
H10602/06	0.313	Die before Treatment	Died	BCL2 rearranged

H16064/06	0.216	Die before Treatment	Died	BCL2 rearranged
H9596/05	0.111	R-CHOP	unknown	BCL6 rearranged
H14627/06	0.093	unknown	unknown	BCL2, BCL6 rearranged

[1]Samples are collected from two clinical trails as well as local cases in HMDS (Haematological Malignancy Diagnostic Service, St. James Hospital, Leeds) Sample ID start with RCH are records from R-CHOP treated trials and start with LY10[10] are records from a dose-modified CODOX-M/IVAC trial. The rest samples are HMDS local patients.

[2]BLprob is the Burkitt lymphoma probability generated by BDC (Burkitt lymphoma and Diffuse large B-cell lymphoma classifier developed in our work).

[3]R-CHOP is usually used to treat DLBCL, named after drugs (Rituximab, Cyclophosphamide, doxorubicin (known as Hydroxydaunomycin) , vincristine (known as Oncovin), Prednisolone); CODOX-M/IVAC is a treatment usually for BL or BL like patients, names after drugs (Cyclophosphamide, vincristine (known as Oncovin), Doxorubicin, Methotrexate, Ifosfamide, Etoposide (known as Vepesid) and Cytarabine (known as Ara-C))

[4]In R-CHOP trails, all patients are evaluated with treatment response and this is used in the study.

Appendix C

Detailed GSEA results

Table 1: Enriched gene sets in *MYC*-rearranged GCB cases

NAME	SIZE	NES	NOM p-val
WALLACE_PROSTATE_CANCER_UP	20	-1.7062083	0
HUMMEL_BURKITTS_LYMPHOMA_UP	39	-1.6270509	0
HOSHIDA_LIVER_CANCER_LATE_RECURRENCE_DN	66	-1.6134428	0.001890359
GSE36476_YOUNG_VS_OLD_DONOR_MEMORY_CD4_TCELL_16H_TSST_ACT_UP	186	-1.5362979	0.009689922
GSE25087_TREG_VS_TCONV_ADULT_DN	134	-1.5174266	0.011049724
AMIT_SERUM_RESPONSE_480_MCF10A	35	-1.5550736	0.011472276
CHEN_NEUROBLASTOMA_COPY_NUMBER_GAINS	42	-1.5980096	0.012072435
LIU_TOPBP1_TARGETS	15	-1.6500998	0.013513514
KYNG_ENVIRONMENTAL_STRESS_RESPONSE_DN	15	-1.6545824	0.013972056
GSE3982_BCELL_VS_BASOPHIL_UP	187	-1.4730427	0.01814516
XU_GH1_EXOGENOUS_TARGETS_UP	62	-1.521727	0.01968504
SCHLOSSER_MYC_AND_SERUM_RESPONSE_SYNERGY	32	-1.5825396	0.022088353
ACOSTA_PROLIFERATION_INDEPENDENT_MYC_TARGETS_UP	78	-1.5817859	0.022774328
BROWN_MYELOID_CELL_DEVELOPMENT_DN	109	-1.48494	0.024193548
REGULATION_OF_MAPKKK_CASCADE	18	-1.5556397	0.02504817

GSE360_L_MAJOR_VS_T_GONDII_MAC_DN	193	-1.5165459	0.026639344
HASLINGER_B_CLL_WITH_13Q14_DELETION	23	-1.490647	0.028957529
GSE15659_CD45RA_NEG_CD4_TCELL_VS_RESTING_TREG_DN	114	-1.4533107	0.030042918
AZARE_NEOPLASTIC_TRANSFORMATION_BY_STAT3_UP	16	-1.4437289	0.031894933
MYC_UP.V1_UP	123	-1.6026452	0.035051547
GSE13306_RA_VS_UNTREATED_TCONV_UP	140	-1.399172	0.035433073
BASSO_CD40_SIGNALING_DN	66	-1.445393	0.03731343
BIOCARTA_CARDIACEGF_PATHWAY	18	-1.5105307	0.039252337
MARIADASON_REGULATED_BY_HISTONE_ACETYLATION_DN	38	-1.4294868	0.04411765
REGULATION_OF_G_PROTEIN_COUPLED_RECEPTOR_PROTEIN_SIGNALING_PATHWAY	23	-1.4829485	0.048076924
REACTOME_3_UTR_MEDIATED_TRANSLATIONAL_REGULATION	89	-1.7070949	0.04918033

Table 2: Enriched gene sets in GSE10846 *MYC* mRNA highly expressed cases (FDR < 0.05)

NAME	SIZE	NES	NOM p-val	FDR q-val
REACTOME_NONSENSE_MEDIATED_DECAY_ENHANCED_BY_THE_EXON_JUNCTION_COMPLEX	92	2.237787	0	0.002958
REACTOME_INFLUENZA_LIFE_CYCLE	120	2.160322	0	0.003334
RIBONUCLEOPROTEIN_COMPLEX	136	2.145989	0	0.003738
CHNG_MULTIPLE_MYELOMA_HYPERPLOID_UP	44	2.160974	0	0.00389
MYC_UP.V1_UP	169	2.162476	0	0.004173
ACOSTA_PROLIFERATION_INDEPENDENT_MYC_TARGETS_UP	77	2.079432	0	0.004571
SCHLOSSER_MYC_TARGETS_AND_SERUM_RESPONSE_UP	43	2.062894	0	0.004675
REACTOME_SRP_DEPENDENT_COTRANSLATIONAL_PROTEIN_TARGETING_TO_MEMBRANE	94	2.059008	0	0.004708
ORGANELLE_ENVELOPE	163	2.065575	0.001953	0.004724
RNA_SPLICING	88	2.079537	0	0.0048
KEGG_AMINOACYL_TRNA_BIOSYNTHESIS	41	2.053966	0	0.004802
GSE15930_NAIVE_VS_48H_IN_VITRO_STIM_IFNAB_CD8_TCELL_DN	188	2.059276	0	0.004813
MITOCHONDRIAL_ENVELOPE	94	2.055216	0.00198	0.004823
PENG_LEUCINE_DEPRIVATION_DN	178	2.063365	0	0.004836
DANG_MYC_TARGETS_UP	129	2.067121	0	0.004869
HELICASE_ACTIVITY	50	2.083328	0	0.004892
ENVELOPE	163	2.065575	0.001953	0.004899
BILANGES_RAPAMYCIN_SENSITIVE_VIA_TSC1_AND_TSC2	67	2.074603	0	0.004944

IRITANI_MAD1_TARGETS_DN	43	2.068284	0	0.004949
NUCLEASE_ACTIVITY	54	2.046343	0	0.004952
YAO_TEMPORAL_RESPONSE_TO_PROGESTERONE_CLUSTER_11	94	2.049338	0	0.005027
RNA_PROCESSING	161	2.068467	0	0.005155
NUCLEOLUS	118	2.085108	0	0.005163
ZHANG_RESPONSE_TO_CANTHARIDIN_DN	65	2.07496	0	0.005169
SCHUHMACHER_MYC_TARGETS_UP	74	2.163868	0	0.005216
REACTOME_INFLUENZA_VIRAL_RNA_TRANSCRIPTION_AND_REPLICATION	87	2.123461	0.00198	0.005269
DANG_REGULATED_BY_MYC_UP	67	2.086646	0	0.005387
ZHAN_VARIABLE_EARLY_DIFFERENTIATION_GENES_DN	30	2.035718	0	0.005677
REACTOME_3_UTR_MEDIATED_TRANSLATIONAL_REGULATION	90	2.086994	0.002058	0.005723
REACTOME_TRNA_AMINOACYLATION	42	2.091108	0	0.005793
KEGG_SELENOAMINO_ACID_METABOLISM	26	2.029383	0	0.006025
SPLICEOSOME	51	2.030412	0	0.006034
GROSS_HYPOXIA_VIA_HIF1A_UP	73	2.107907	0	0.006074
MITOCHONDRIAL_MEMBRANE	83	2.091245	0.001931	0.006206
REACTOME_METABOLISM_OF_MRNA	191	2.110272	0	0.006329
YAO_TEMPORAL_RESPONSE_TO_PROGESTERONE_CLUSTER_14	138	2.09189	0	0.006684
BILANGES_SERUM_AND_RAPAMYCIN_SENSITIVE_GENES	54	2.095068	0	0.006782
KEGG_SPLICEOSOME	113	2.171054	0	0.006955

GSE24026_PD1_LIGATION_VS_CTRL_IN_ACT_TCELL_LINE_DN	189	2.007988	0.001898	0.00734
SANSOM_APC_MYC_TARGETS	198	2.003705	0	0.007381
REACTOME_MITOCHONDRIAL_TRNA_AMINOACYLATION	21	2.002501	0	0.007391
MITOCHONDRIAL_TRANSPORT	21	2.00871	0	0.007404
KARLSSON_TGFB1_TARGETS_UP	118	2.004638	0	0.007409
KIM_MYC_AMPLIFICATION_TARGETS_UP	186	2.006415	0	0.007423
SCHLOSSER_MYC_TARGETS_AND_SERUM_RESPONSE_DN	45	2.015008	0	0.00745
STRUCTURE_SPECIFIC_DNA_BINDING	51	2.010603	0	0.007496
DNA_HELICASE_ACTIVITY	24	2.010932	0	0.007599
MITOCHONDRIAL_PART	137	2.012392	0.001984	0.007681
JAIN_NFKB_SIGNALING	71	1.991842	0	0.008309
MRNA_METABOLIC_PROCESS	78	1.987745	0	0.008509
GSE15930_NAIVE_VS_48H_IN_VITRO_STIM_CD8_TCELL_DN	189	1.988084	0	0.008619
REACTOME_TRANSLATION	128	2.174885	0	0.009114
SANSOM_APC_TARGETS_REQUIRE_MYC	186	1.981014	0	0.009262
PID_MYC_ACTIVPATHWAY	76	1.980089	0	0.009262
WELCSH_BRCA1_TARGETS_DN	137	1.978041	0.003883	0.009306
GSE22886_UNSTIM_VS_IL2_STIM_NKCELL_DN	189	1.975732	0	0.009475
REACTOME_ACTIVATION_OF_THE_MRNA_UPON_BINDING_OF_THE_CAP_BINDING_COMPLEX_AND_EIFS_AND_SUBSEQUENT_BINDING_TO_43S	51	1.976206	0.006073	0.009545
RIBONUCLEOPROTEIN_COMPLEX_BIOGENESIS_AND_ASSEMBLY	80	1.971054	0	0.009957

MITOCHONDRIAL_INNER_MEMBRANE	64	1.97122	0	0.010089
TIEN_INTESTINE_PROBIOTICS_6HR_UP	49	1.964846	0.005871	0.010937
MRNA_PROCESSING_GO_0006397	68	1.962363	0	0.011071
GSE31082_DN_VS_CD8_SP_THYMOCYTE_UP	184	1.955439	0	0.012269
REACTOME_TRANSCRIPTION	157	1.95321	0.001894	0.012452
REACTOME_PROCESSING_OF_CAPPED_INTRON_CONTAINING_PRE_MRNA	128	1.952255	0.003759	0.012459
REACTOME_MRNA_SPLICING_MINOR_PATHWAY	40	1.949148	0	0.012853
SCHLOSSER_MYC_AND_SERUM_RESPONSE_SYNERGY	31	1.944831	0.001901	0.013061
LI_DCP2_BOUND_MRNA	84	1.945249	0	0.013188
ORGANELLE_INNER_MEMBRANE	72	1.946149	0	0.013242
REACTOME_RNA_POL_III_TRANSCRIPTION	33	1.942189	0	0.013406
HEDENFALK_BREAST_CANCER_HEREDITARY_VS_SPORADIC	45	1.938536	0	0.01396
MENSSEN_MYC_TARGETS	46	1.937248	0	0.014037
GSE31082_DN_VS_DP_THYMOCYTE_UP	185	1.93902	0	0.014083
SCHLOSSER_MYC_TARGETS_REPRESSED_BY_SERUM	156	1.933287	0	0.01466
TRANSCRIPTION_FROM_RNA_POLYMERASE_III_PROMOTER	19	1.932149	0	0.014819
GRADE_METASTASIS_DN	42	1.928608	0.007828	0.015434
REACTOME_CYTOSOLIC_TRNA_AMINOACYLATION	24	1.919238	0	0.01712
REACTOME_RNA_POL_III_TRANSCRIPTION_INITIATION_FROM_TYPE_2_PROMOTER	23	1.919735	0	0.017178
LEE_METASTASIS_AND_RNA_PROCESSING_UP	17	1.909566	0	0.017389

RNA_HELICASE_ACTIVITY	24	1.908508	0	0.017487
MARCINIAK_ER_STRESS_RESPONSE_VIA_CHOP	24	1.910643	0.002092	0.017573
PID_P53REGULATIONPATHWAY	57	1.911712	0.003968	0.017585
IVANOVA_HEMATOPOIESIS_INTERMEDIATE_PROGENITOR	140	1.909616	0	0.017588
GARCIA_TARGETS_OF_FLI1_AND_DAX1_DN	161	1.907069	0	0.017631
ZAMORA_NOS2_TARGETS_UP	63	1.912042	0.003868	0.017744
GSE24634_NAIVE_CD4_TCELL_VS_DAY5_IL4_CONV_TREG_DN	193	1.907248	0	0.017766
GSE28237_FOLLICULAR_VS_EARLY_GC_BCELL_DN	189	1.915134	0	0.017874
REACTOME_DEADENYLATION_DEPENDENT_MRNA_DECAY	42	1.91267	0	0.017908
REACTOME_MRNA_PROCESSING	148	1.91381	0.007634	0.017966
REACTOME_RNA_POL_III_TRANSCRIPTION_INITIATION_FROM_TYPE_3_PROMOTER	26	1.904379	0	0.018174
REACTOME_PEPTIDE_CHAIN_ELONGATION	71	1.902682	0.01222	0.018604
YAO_TEMPORAL_RESPONSE_TO_PROGESTERONE_CLUSTER_17	173	1.897515	0.005859	0.019176
KEGG_PURINE_METABOLISM	147	1.897948	0	0.019311
XU_RESPONSE_TO_TRETINOIN_AND_NSC682994_DN	15	1.895101	0	0.019331
REACTOME_MRNA_SPLICING	99	1.898161	0.007678	0.019447
REACTOME_PYRUVATE_METABOLISM_AND_CITRIC_ACID_TCA_CYCLE	40	1.895354	0	0.019455
GSE31082_DN_VS_CD4_SP_THYMOCYTE_UP	180	1.892536	0	0.019534
COLLER_MYC_TARGETS_UP	24	1.89273	0	0.019683
MITOCHONDRION_ORGANIZATION_AND_BIOGENESIS	45	1.887347	0.001931	0.020502

GSE24634_TREG_VS_TCONV_POST_DAY3_IL4_CONVERSION_UP	193	1.888615	0.003817	0.02055
STRUCTURAL_CONSTITUENT_OF_RIBOSOME	68	1.887546	0.001992	0.020655
GSE15930_NAIVE_VS_48H_IN_VITRO_STIM_IL12_CD8_TCELL_DN	189	1.885251	0	0.020937
JUBAN_TARGETS_OF_SPI1_AND_FLI1_DN	81	1.881741	0	0.021554
GSE24634_TREG_VS_TCONV_POST_DAY5_IL4_CONVERSION_UP	191	1.882035	0.001883	0.021707
DEN_INTERACT_WITH_LCA5	23	1.877354	0.003922	0.022614
GSE22886_UNSTIM_VS_IL15_STIM_NKCELL_DN	192	1.875597	0	0.022857
ENDONUCLEASE_ACTIVITY	24	1.873231	0	0.023298
LUI_THYROID_CANCER_CLUSTER_3	24	1.871209	0.003945	0.023371
PROTEIN_RNA_COMPLEX_ASSEMBLY	63	1.869007	0	0.023463
GSE15930_NAIVE_VS_24H_IN_VITRO_STIM_CD8_TCELL_DN	191	1.871633	0	0.023501
BIOCARTA_ARF_PATHWAY	17	1.869067	0	0.023631
TRANSLATION	167	1.869787	0	0.023667
GSE17721_POLYIC_VS_GARDIQUIMOD_8H_BMDM_DN	175	1.871718	0.002004	0.023683
HOSHIDA_LIVER_CANCER_SUBCLASS_S2	111	1.861833	0.002045	0.024009
RPS14_DN.V1_DN	179	1.862126	0.001984	0.024095
POMEROY_MEDULLOBLASTOMA_PROGNOSIS_DN	42	1.86342	0.007797	0.024219
KEGG_RNA_DEGRADATION	57	1.862283	0.001953	0.024265
CAIRO_PML_TARGETS_BOUND_BY_MYC_UP	23	1.853844	0	0.024288
REACTOME_RNA_POL_I_RNA_POL_III_AND_MITOCHONDRIAL_TRANSCRIPTION	80	1.854471	0	0.024294

DEOXYRIBONUCLEASE_ACTIVITY	22	1.854751	0.001988	0.024398
REACTOME_DNA_REPAIR	105	1.863515	0.003861	0.0244
GSE7764_IL15_TREATED_VS_CTRL_NK_CELL_24H_UP	186	1.865385	0	0.02442
RESPONSE_TO_DNA_DAMAGE_STIMULUS	156	1.862514	0.00578	0.024421
PID_MYC_PATHWAY	25	1.855168	0.001969	0.024474
ATP_DEPENDENT_HELICASE_ACTIVITY	26	1.856201	0.013436	0.024494
REACTOME_RNA_POL_I_TRANSCRIPTION_TERMINATION	21	1.855471	0	0.024548
PRAMOONJAGO_SOX4_TARGETS_DN	48	1.86357	0.003831	0.024596
WIERENGA_PML_INTERACTOME	39	1.86419	0.004082	0.024657
GSE3982_NEUTROPHIL_VS_BCELL_DN	191	1.856888	0	0.024661
BIOCARTA_EIF_PATHWAY	15	1.856208	0.001969	0.024687
REACTOME_PURINE_METABOLISM	32	1.851566	0.009728	0.024761
TRANSCRIPTION_INITIATION	35	1.859122	0.001949	0.024802
REACTOME_ZINC_TRANSPORTERS	15	1.856996	0.003922	0.024848
KEEN_RESPONSE_TO_ROSIGLITAZONE_UP	36	1.857252	0.001957	0.024944
PROTEIN_FOLDING	58	1.857888	0.002	0.024956
REACTOME_RNA_POL_III_TRANSCRIPTION_TERMINATION	19	1.845712	0	0.025197
OUTER_MEMBRANE	25	1.845789	0.014113	0.025352
RNA_SPLICINGVIA_TRANSESTERIFICATION_REACTIONS	34	1.843808	0.001961	0.025403
GSE17974_CTRL_VS_ACT_IL4_AND_ANTI_IL12_12H_CD4_TCELL_DN	187	1.844213	0	0.025448

MITOCHONDRIAL_OUTER_MEMBRANE	18	1.848651	0.015968	0.025467
GSE24634_TEFF_VS_TCONV_DAY5_IN_CULTURE_UP	194	1.845873	0.001883	0.025482
VANTVEER_BREAST_CANCER_BRCA1_UP	33	1.847531	0	0.025502
SINGLE_STRANDED_DNA_BINDING	32	1.846921	0.001976	0.025512
REACTOME_LATE_PHASE_OF_HIV_LIFE_CYCLE	96	1.846056	0.003831	0.025599
WALLACE_PROSTATE_CANCER_UP	20	1.847743	0	0.025633
GSE17974_0H_VS_24H_IN_VITRO_ACT_CD4_TCELL_DN	182	1.840415	0	0.026351
GSE22886_UNSTIM_VS_STIM_MEMORY_TCELL_DN	193	1.840811	0.005725	0.026421
CHAUHAN_RESPONSE_TO_METHOXYESTRADIOL_UP	48	1.838625	0.001931	0.026841
RIBOSOME_BIOGENESIS_AND_ASSEMBLY	16	1.836711	0	0.027355
TRANSLATION_INITIATION_FACTOR_ACTIVITY	23	1.835631	0	0.027639
DNA_DAMAGE_RESPONSESIGNAL_TRANSDUCTION	34	1.831681	0.004	0.028744
REACTOME_RNA_POL_III_CHAIN_ELONGATION	17	1.828878	0	0.02886
MALONEY_RESPONSE_TO_17AAG_DN	73	1.831707	0.003883	0.028931
KEGG_HUNTINGTONS_DISEASE	166	1.829137	0.011696	0.028993
REACTOME_FORMATION_OF_RNA_POL_II_ELONGATION_COMPLEX_	39	1.828043	0	0.028997
CHANDRAN_METASTASIS_TOP50_UP	35	1.830293	0.013861	0.029042
NUCLEOLAR_PART	18	1.82951	0	0.029065
REACTOME_CITRIC_ACID_CYCLE_TCA_CYCLE	19	1.82544	0.007937	0.029186
DOUBLE_STRANDED_DNA_BINDING	28	1.825483	0	0.029341

GSE22886_NAIVE_CD4_TCELL_VS_12H_ACT_TH1_DN	193	1.826291	0.003831	0.029404
REACTOME_BRANCHED_CHAIN_AMINO_ACID_CATABOLISM	17	1.825746	0.001869	0.029423
REACTOME_RNA_POL_II_TRANSCRIPTION	93	1.820666	0.007813	0.030811
REACTOME_DESTABILIZATION_OF_MRNA_BY_TRISTETRAPROLIN_TTP	17	1.818059	0.001821	0.031801
REACTOME_FORMATION_OF_THE_TERNARY_COMPLEX_AND_SUBSEQUENTLY_THE_43S_COMPLEX	44	1.816298	0.012146	0.031944
RAHMAN_TP53_TARGETS_PHOSPHORYLATED	21	1.811845	0.009615	0.031969
GSE19825_NAIVE_VS_IL2RAHIGH_DAY3_EFF_CD8_TCELL_DN	187	1.815869	0.003839	0.031999
NUCLEAR_BODY	30	1.816852	0.005725	0.032073
REACTOME_REGULATORY_RNA_PATHWAYS	20	1.816419	0	0.032098
GSE17974_CTRL_VS_ACT_IL4_AND_ANTI_IL12_24H_CD4_TCELL_DN	185	1.811886	0	0.032143
PORE_COMPLEX	34	1.814131	0.007421	0.032289
BIOCARTA_TEL_PATHWAY	18	1.811966	0.008247	0.032312
NUCLEOTIDYLTRANSFERASE_ACTIVITY	46	1.813293	0.003914	0.032348
GSE15930_NAIVE_VS_24H_IN_VITRO_STIM_IL12_CD8_TCELL_DN	192	1.814205	0.001905	0.032453
REACTOME_AMINO_ACID_SYNTHESIS_AND_INTERCONVERSION_TRANSAMINATION	16	1.812016	0	0.032494
GSE27786_LIN_NEG_VS_NKCELL_UP	186	1.812082	0	0.03266
GSE27786_LIN_NEG_VS_MONO_MAC_UP	187	1.809192	0	0.032784
UDAYAKUMAR_MED1_TARGETS_UP	131	1.805964	0	0.033851
DNA_DEPENDENT_ATPASE_ACTIVITY	21	1.806346	0	0.03389
DNA_REPAIR	120	1.804955	0.013725	0.034067

DAIRKEE_CANCER_PRONE_RESPONSE_BPA	49	1.803839	0.005769	0.034337
EXONUCLEASE_ACTIVITY	19	1.802537	0	0.034429
BROWN_MYELOID_CELL_DEVELOPMENT_DN	119	1.800799	0.001923	0.034462
RHODES_CANCER_META_SIGNATURE	64	1.800457	0.015009	0.034469
GENERAL_RNA_POLYMERASE_II_TRANSCRIPTION_FACTOR_ACTIVITY	31	1.802761	0	0.034546
RNA_EXPORT_FROM_NUCLEUS	20	1.801665	0	0.034581
MITOCHONDRIAL_MEMBRANE_PART	50	1.800833	0.007692	0.034616
TRNA_METABOLIC_PROCESS	19	1.800999	0.009381	0.034723
BHATTACHARYA_EMBRYONIC_STEM_CELL	83	1.797841	0.001984	0.03484
REACTOME_MITOCHONDRIAL_PROTEIN_IMPORT	47	1.799264	0	0.034843
MEIOTIC_RECOMBINATION	16	1.798716	0.007648	0.034858
GSE17721_LPS_VS_PAM3CSK4_4H_BMDM_DN	183	1.797941	0.001953	0.034997
KEGG_LYSINE_DEGRADATION	44	1.79478	0.003831	0.035032
GSE32423_MEMORY_VS_NAIVE_CD8_TCELL_IL7_DN	179	1.794899	0.001957	0.035144
RESPONSE_TO_ENDOGENOUS_STIMULUS	192	1.795292	0.006	0.035155
GSE17721_POLYIC_VS_PAM3CSK4_6H_BMDM_DN	186	1.794309	0.00188	0.035159
GSE27786_NEUTROPHIL_VS_MONO_MAC_DN	184	1.793777	0	0.035183
GSE18791_UNSTIM_VS_NEWCATSLE_VIRUS_DC_10H_UP	185	1.795507	0.001969	0.035256
DNA_DIRECTED_RNA_POLYMERASEII_HOLOENZYME	64	1.795765	0.003968	0.035351
GSE15930_NAIVE_VS_24H_IN_VITRO_STIM_INFAB_CD8_TCELL_DN	190	1.796217	0.003774	0.035375

BOYAULT_LIVER_CANCER_SUBCLASS_G3_UP	182	1.79228	0.003929	0.035628
REACTOME_METABOLISM_OF_NON_CODING_RNA	46	1.791416	0	0.035694
TRANSLATION_REGULATOR_ACTIVITY	38	1.791824	0.001938	0.035727
ORGANELLE_OUTER_MEMBRANE	24	1.790667	0.02008	0.035852
TRANSLATION_FACTOR_ACTIVITY_NUCLEIC_ACID_BINDING	36	1.789821	0.003914	0.036075
NUCLEAR_MEMBRANE	48	1.788045	0.003861	0.036105
ELVIDGE_HYPOXIA_DN	141	1.788825	0.023166	0.036112
PROTEIN_N_TERMINUS_BINDING	38	1.789087	0.00396	0.036179
MULLIGAN_NTF3_SIGNALING_VIA_INSR_AND_IGF1R_UP	20	1.788197	0.005758	0.036195
GSE9006_HEALTHY_VS_TYPE_1_DIABETES_PBMC_1MONTH_POST_DX_UP	194	1.787545	0.007874	0.036201
KEGG_RNA_POLYMERASE	27	1.78679	0.001916	0.036361
COLLIS_PRKDC_SUBSTRATES	20	1.786065	0.001942	0.036391
KEGG_RIBOSOME	73	1.786309	0.014523	0.036453
REACTOME_TRANSPORT_OF_MATURE_TRANSCRIPT_TO_CYTOPLASM	49	1.783152	0.017375	0.037562
GSE11924_TFH_VS_TH1_CD4_TCELL_DN	188	1.782776	0	0.037588
GSE17721_PAM3CSK4_VS_GADIQUIMOD_6H_BMDM_UP	189	1.779248	0.001938	0.038717
DNA_CATABOLIC_PROCESS	23	1.779351	0.001984	0.038813
BURTON_ADIPOGENESIS_PEAK_AT_16HR	40	1.779919	0.01476	0.038847
CHNG_MULTIPLE_MYELOMA_HYPERPLOID_DN	28	1.779506	0.012448	0.038881
RIBOSOME	36	1.777931	0.001942	0.0389

HEDENFALK_BREAST_CANCER_BRACX_DN	20	1.778482	0.015414	0.038907
REACTOME_RNA_POL_II_PRE_TRANSCRIPTION_EVENTS	54	1.776734	0.003929	0.039342
GSE10239_KLRG1INT_VS_KLRG1HIGH_EFF_CD8_TCELL_UP	182	1.775342	0.005871	0.039685
NUCLEAR_ENVELOPE	70	1.775626	0.001946	0.039697
FLECHNER_BIOPSY_KIDNEY_TRANSPLANT_OK_VS_DONOR_DN	25	1.772921	0.013566	0.040594
MRNA_BINDING	20	1.772002	0.015779	0.040662
RIBONUCLEASE_ACTIVITY	24	1.771337	0	0.040677
E2F1_UP.V1_UP	182	1.772216	0.007782	0.040787
DNA_REPLICATION	98	1.771416	0.01165	0.040814
REACTOME_ELONGATION_ARREST_AND_RECOVERY	27	1.769001	0.009488	0.041083
TRANSLATIONAL_INITIATION	36	1.770249	0.00198	0.041099
ATP_DEPENDENT_RNA_HELICASE_ACTIVITY	17	1.769108	0.007648	0.041199
REACTOME_TRANSCRIPTION_COUPLED_NER_TC_NER	44	1.768321	0.013619	0.041207
GSE27786_LSK_VS_NKTCELL_UP	182	1.769402	0	0.041254
REACTOME_HIV_LIFE_CYCLE	109	1.766241	0.015066	0.04187
REACTOME_NUCLEOTIDE_EXCISION_REPAIR	49	1.766385	0.015504	0.042002
TRANSCRIPTION_INITIATION_FROM_RNA_POLYMERASE_II_PROMOTER	29	1.765355	0.001969	0.042146
REACTOME_DESTABILIZATION_OF_MRNA_BY_KSRP	17	1.764539	0.013258	0.042353
DNA_RECOMBINATION	46	1.763909	0.01006	0.042507
BILD_MYC_ONCOGENIC_SIGNATURE	192	1.761332	0.00381	0.043532

KEGG_CITRATE_CYCLE_TCA_CYCLE	30	1.759984	0.013752	0.043927
REACTOME_INTERACTIONS_OF_VPR_WITH_HOST_CELLULAR_PROTEINS	32	1.760191	0.005747	0.044009
SMALL_NUCLEAR_RIBONUCLEOPROTEIN_COMPLEX	22	1.759397	0.011299	0.044074
REACTOME_REGULATION_OF_GLUCOKINASE_BY_GLUCOKINASE_REGULATORY_PROTEIN	27	1.757871	0.003861	0.044543
GSE17721_LPS_VS_CPG_1H_BMDM_UP	189	1.757084	0	0.044642
NUNODA_RESPONSE_TO_DASATINIB_IMATINIB_UP	29	1.757406	0.007767	0.044682
GSE27786_BCELL_VS_NEUTROPHIL_UP	188	1.757936	0.007707	0.044687
KEGG_ALANINE_ASPARTATE_AND_GLUTAMATE_METABOLISM	32	1.753416	0.005906	0.044751
LI_WILMS_TUMOR_VS_FETAL_KIDNEY_1_DN	156	1.752443	0.026769	0.04486
STARK_HYPPOCAMPUS_22Q11_DELETION_DN	17	1.753514	0.018219	0.044888
REACTOME_FANCONI_ANEMIA_PATHWAY	21	1.75397	0.001942	0.044912
ACETYLTRANSFERASE_ACTIVITY	25	1.752624	0.005671	0.044961
CHEMNITZ_RESPONSE_TO_PROSTAGLANDIN_E2_UP	134	1.75516	0.022514	0.044988
GSE17974_CTRL_VS_ACT_IL4_AND_ANTI_IL12_4H_CD4_TCELL_DN	183	1.754135	0.003891	0.044991
KEGG_PYRIMIDINE_METABOLISM	89	1.756109	0.01004	0.045017
PROTEIN_UBIQUITINATION	39	1.755366	0.005929	0.04506
REACTOME_TRANSPORT_OF_MATURE_MRNA_DERIVED_FROM_AN_INTRONLESS_TRANSCRIPT	31	1.751654	0.001931	0.045085
KEGG_ARGININE_AND_PROLINE_METABOLISM	50	1.754138	0.013944	0.04517
TAKAO_RESPONSE_TO_UVB_RADIATION_DN	96	1.755535	0.001969	0.045171
HESS_TARGETS_OF_HOXA9_AND_MEIS1_UP	61	1.754418	0.00578	0.045234

REACTOME_RNA_POL_I_TRANSCRIPTION_INITIATION	24	1.751065	0	0.045254
BASSO_B_LYMPHOCYTE_NETWORK	137	1.749485	0.001949	0.045871
REACTOME_ACTIVATION_OF_BH3_ONLY_PROTEINS	15	1.749073	0.00759	0.045928
MORI_EMU_MYC_LYMPHOMA_BY_ONSET_TIME_UP	97	1.747861	0.005929	0.046219
REACTOME_FORMATION_OF_TRANSCRIPTION_COUPLED_NER_TC_NER_REPAIR_COMPLEX	29	1.748061	0.003861	0.04628
GSE22886_NAIVE_CD4_TCELL_VS_48H_ACT_TH1_DN	191	1.745037	0.015686	0.046482
REACTOME_MICRORNA_MIRNA_BIOGENESIS	17	1.74529	0.001931	0.046505
GSE22886_NAIVE_TCELL_VS_NEUTROPHIL_UP	190	1.746549	0.032882	0.046542
DNA_INTEGRITY_CHECKPOINT	24	1.746737	0.005894	0.046591
PID_FANCONI_PATHWAY	46	1.746106	0.009709	0.046632
RNA_DEPENDENT_ATPASE_ACTIVITY	18	1.745355	0.022989	0.046642
PID_P73PATHWAY	76	1.74572	0.005736	0.046665
REACTOME_DESTABILIZATION_OF_MRNA_BY_BRF1	17	1.744316	0.011029	0.04669
REACTOME_NEP_NS2_INTERACTS_WITH_THE_CELLULAR_EXPORT_MACHINERY	27	1.743173	0.003831	0.046716
DNA_DAMAGE_CHECKPOINT	20	1.743963	0.006073	0.046734
GSE1460_INTRATHYMIC_T_PROGENITOR_VS_NAIVE_CD4_TCELL_CORD_BLOOD_UP	190	1.743274	0.013861	0.046847
TRANSFERASE_ACTIVITY_TRANSFERRING_GROUPS_OTHER_THAN_AMINO_ACYL_GROUPS	47	1.741538	0.00578	0.04693
VEGF_A_UP.V1_DN	190	1.741566	0.002053	0.047085
NUCLEOBASENUCLEOSIDE_AND_NUCLEOTIDE_METABOLIC_PROCESS	49	1.742225	0.001927	0.047101
GSE36392_TYPE_2_MYELOID_VS_MAC_IL25_TREATED_LUNG_DN	183	1.741662	0.00789	0.047182

MITOCHONDRIAL_LUMEN	44	1.738406	0.007968	0.047855
MITOCHONDRIAL_MATRIX	44	1.738406	0.007968	0.048026
REACTOME_FORMATION_OF_THE_HIV1_EARLY_ELONGATION_COMPLEX	29	1.738688	0.001919	0.04804
KEGG_NUCLEOTIDE_EXCISION_REPAIR	44	1.736138	0.017442	0.048048
LIGASE_ACTIVITY	95	1.736347	0.013672	0.048109
LY_AGING_OLD_DN	55	1.737296	0.013514	0.048124
PID_BARD1PATHWAY	29	1.737533	0.028681	0.048161
GSE17721_POLYIC_VS_PAM3CSK4_8H_BMDM_DN	185	1.738735	0	0.048195
WANG_TARGETS_OF_MLL_CBP_FUSION_DN	45	1.736882	0.005929	0.048205
REACTOME_SYNTHESIS_OF_GLYCOSYLPHOSPHATIDYLINOSITOL_GPI	16	1.736434	0.006148	0.048253
ABRAMSON_INTERACT_WITH_AIRE	43	1.735281	0.003899	0.048324
GSE17721_PAM3CSK4_VS_GADIQUIMOD_8H_BMDM_UP	188	1.734982	0.001949	0.048327
REACTOME_TCA_CYCLE_AND_RESPIRATORY_ELECTRON_TRANSPORT	117	1.733323	0.029183	0.048913
GSE19825_CD24LOW_VS_IL2RA_HIGH_DAY3_EFF_CD8_TCELL_DN	191	1.733518	0.001969	0.048997
REACTOME_PREFOLDIN_MEDIATED_TRANSFER_OF_SUBSTRATE_TO_CCT_TRIC	24	1.731846	0.013645	0.049238
PUJANA_BRCA_CENTERED_NETWORK	117	1.731513	0.017308	0.049287
GSE24634_NAIVE_CD4_TCELL_VS_DAY3_IL4_CONV_TREG_DN	189	1.731916	0.013384	0.049339
REACTOME_RESOLUTION_OF_AP_SITES_VIA_THE_MULTIPLE_NUCLEOTIDE_PATCH_REPLACEMENT_PATHWAY	17	1.732164	0.007859	0.049352

Table 3: Enriched gene sets in Monti dataset MYC mRNA highly expressed cases (FDR < 0.05 and NES > 2)

NAME	SIZE	NES	NOMp-val	FDR q-val
TIEN_INTESTINE_PROBIOTICS_6HR_UP	55	2.302072	0	2.77E-04
RIBONUCLEOPROTEIN_COMPLEX_BIOGENESIS_AND_ASSEMBLY	84	2.303925	0	3.69E-04
DANG_MYC_TARGETS_UP	139	2.308868	0	5.54E-04
RNA_PROCESSING	170	2.280573	0	8.87E-04
HELICASE_ACTIVITY	51	2.28162	0	0.00106486
JUBAN_TARGETS_OF_SPI1_AND_FLI1_DN	87	2.311738	0	0.0011072
NUCLEOLUS	122	2.225878	0	0.00117365
RNA_HELICASE_ACTIVITY	24	2.226765	0	0.00125749
REACTOME_INFLUENZA_LIFE_CYCLE	134	2.251753	0	0.00126821
GROSS_HYPOXIA_VIA_HIF1A_UP	75	2.215831	0	0.00127863
RIBONUCLEOPROTEIN_COMPLEX	142	2.243005	0	0.00130454
SCHUHMACHER_MYC_TARGETS_UP	77	2.22853	0	0.00135422
ATP_DEPENDENT_HELICASE_ACTIVITY	27	2.25236	0	0.00142674
REACTOME_DEADENYLATION_DEPENDENT_MRNA_DECAY	44	2.243377	0	0.00143499
GSE17721_LPS_VS_POLYIC_12H_BMDM_UP	194	2.25802	0	0.00143668
GSE24026_PD1_LIGATION_VS_CTRL_IN_ACT_TCELL_LINE_DN	198	2.229192	0	0.00146707
TRANSLATION_FACTOR_ACTIVITY_NUCLEIC_ACID_BINDING	38	2.190151	0	0.00176779

SCHLOSSER_MYC_TARGETS_AND_SERUM_RESPONSE_UP	46	2.191599	0	0.00186083
PENG_LEUCINE_DEPRIVATION_DN	184	2.200531	0	0.00190364
TRANSLATION_REGULATOR_ACTIVITY	40	2.185477	0	0.00192369
YAO_TEMPORAL_RESPONSE_TO_PROGESTERONE_CLUSTER_14	141	2.192001	0	0.00196421
PUIFFE_INVASION_INHIBITED_BY_ASCITES_UP	82	2.169003	0	0.00200452
GSE22886_NAIVE_TCELL_VS_NEUTROPHIL_UP	196	2.179475	0	0.00202115
RNA_SPLICING	90	2.182884	0	0.00202202
KEGG_SPLICEOSOME	124	2.17045	0	0.00207876
REACTOME_TRNA_AMINOACYLATION	42	2.174717	0	0.00213166
GSE3982_NEUTROPHIL_VS_TH2_DN	196	2.160707	0	0.00213282
BIOPOLYMER_CATABOLIC_PROCESS	115	2.162733	0	0.00213549
ELVIDGE_HYPOXIA_DN	143	2.171065	0	0.00215872
GSE9006_HEALTHY_VS_TYPE_1_DIABETES_PBMC_1MONTH_POST_DX_UP	197	2.174053	0	0.00216525
PID_MYC_ACTIVPATHWAY	78	2.154827	0	0.00243964
GSE27786_LIN_NEG_VS_MONO_MAC_UP	193	2.137849	0	0.00248582
GSE3982_NEUTROPHIL_VS_TH1_DN	198	2.141705	0	0.00251881
MORI_EMU_MYC_LYMPHOMA_BY_ONSET_TIME_UP	102	2.138363	0	0.002553
FERRANDO_T_ALL_WITH_MLL_ENL_FUSION_DN	87	2.142505	0	0.00259078
GSE19825_NAIVE_VS_IL2RAHIGH_DAY3_EFF_CD8_TCELL_DN	195	2.145963	0	0.00263316
DAIRKEE_CANCER_PRONE_RESPONSE_BPA	51	2.145384	0	0.00263938

MRNA_METABOLIC_PROCESS	82	2.13498	0	0.0026413
SCHLOSSER_MYC_TARGETS_REPRESSED_BY_SERUM	157	2.125419	0	0.00266262
ZHANG_RESPONSE_TO_CANTHARIDIN_DN	67	2.126893	0	0.00266663
GSE9006_TYPE_1_VS_TYPE_2_DIABETES_PBMC_AT_DX_UP	194	2.136203	0	0.00268497
GSE9006_HEALTHY_VS_TYPE_2_DIABETES_PBMC_AT_DX_UP	192	2.127727	0	0.00269992
KARLSSON_TGFB1_TARGETS_UP	123	2.146677	0	0.00271545
GSE10239_MEMORY_VS_DAY4.5_EFF_CD8_TCELL_DN	191	2.128472	0	0.0027642
MENSSEN_MYC_TARGETS	52	2.1291	0	0.00283162
PROTEIN_RNA_COMPLEX_ASSEMBLY	65	2.118379	0	0.0029883
ATP_DEPENDENT_RNA_HELICASE_ACTIVITY	17	2.121339	0	0.00300711
BILD_MYC_ONCOGENIC_SIGNATURE	199	2.116307	0	0.00300725
KAUFFMANN_DNA_REPLICATION_GENES	146	2.118818	0	0.00303113
MRNA_PROCESSING_GO_0006397	72	2.113936	0	0.00310077
REACTOME_MRNA_PROCESSING	156	2.109227	0	0.00313122
GSE3982_MAST_CELL_VS_NEUTROPHIL_UP	197	2.10813	0	0.00315184
WATANABE_RECTAL_CANCER_RADIOTHERAPY_RESPONSIVE_UP	108	2.109396	0	0.00317221
DANG_REGULATED_BY_MYC_UP	70	2.10243	0	0.0031761
RNA_DEPENDENT_ATPASE_ACTIVITY	18	2.106267	0	0.00318437
WELCSH_BRCA1_TARGETS_DN	139	2.109618	0	0.00320598
GSE27786_BCELL_VS_ERYTHROBLAST_UP	189	2.102556	0	0.00323086

DNA_HELICASE_ACTIVITY	25	2.099948	0	0.00324565
CAMP_UP.V1_UP	191	2.103629	0	0.00325672
GINESTIER_BREAST_CANCER_20Q13_AMPLIFICATION_DN	171	2.103633	0	0.00331488
BILANGES_RAPAMYCIN_SENSITIVE_VIA_TSC1_AND_TSC2	72	2.092716	0	0.00346915
KEGG_AMINOACYL_TRNA_BIOSYNTHESIS	41	2.091494	0.002016	0.00348747
GSE9006_TYPE_1_DIABETES_AT_DX_VS_1MONTH_POST_DX_PBMC_UP	198	2.095311	0	0.00348825
GSE17721_POLYIC_VS_GARDIQUIMOD_6H_BMDM_DN	194	2.095713	0	0.00349736
PRAMOONJAGO_SOX4_TARGETS_DN	50	2.091705	0	0.00351159
GSE17721_PAM3CSK4_VS_CPG_4H_BMDM_UP	192	2.092802	0	0.00352093
GSE27786_NKTCELL_VS_MONO_MAC_UP	198	2.093667	0	0.00354427
CHAUHAN_RESPONSE_TO_METHOXYESTRADIOL_DN	101	2.093209	0.003929	0.00354543
RIZ_ERYTHROID_DIFFERENTIATION	77	2.092872	0	0.00357428
GSE17721_CTRL_VS_POLYIC_12H_BMDM_UP	189	2.094071	0	0.00360053
MITOCHONDRIAL_PART	140	2.084282	0	0.00375324
JAIN_NFKB_SIGNALING	74	2.085336	0	0.00376023
SCHLOSSER_MYC_TARGETS_AND_SERUM_RESPONSE_DN	46	2.085984	0	0.00378064
ZAMORA_NOS2_TARGETS_UP	66	2.083162	0	0.00380614
ORGANELLE_ENVELOPE	165	2.081863	0	0.00382854
RAHMAN_TP53_TARGETS_PHOSPHORYLATED	21	2.080515	0	0.0038438
IRITANI_MAD1_TARGETS_DN	44	2.07991	0	0.00386929

ENVELOPE	165	2.081863	0	0.00387959
GSE17974_0H_VS_6H_IN_VITRO_ACT_CD4_TCELL_DN	196	2.077659	0	0.00394544
GSE15930_NAIVE_VS_48H_IN_VITRO_STIM_IFNAB_CD8_TCELL_DN	195	2.078481	0	0.00394981
BASSO_B_LYMPHOCYTE_NETWORK	140	2.074359	0	0.00410172
REACTOME_PROCESSING_OF_CAPPED_INTRON_CONTAINING_PRE_MRNA	136	2.073247	0	0.00411165
GSE17974_CTRL_VS_ACT_IL4_AND_ANTI_IL12_4H_CD4_TCELL_DN	189	2.072162	0	0.00411224
HEDENFALK_BREAST_CANCER_BRCA1_VS_BRCA2	160	2.074707	0	0.00413097
RIBOSOME_BIOGENESIS_AND_ASSEMBLY	18	2.069586	0	0.0041316
GSE3982_EOSINOPHIL_VS_TH2_DN	194	2.068422	0	0.0041528
REACTOME_MRNA_SPLICING	107	2.06961	0	0.00417964
RAMASWAMY_METASTASIS_UP	65	2.070026	0	0.00420678
HEDENFALK_BREAST_CANCER_HEREDITARY_VS_SPORADIC	49	2.066296	0	0.00429253
GSE24634_TREG_VS_TCONV_POST_DAY5_IL4_CONVERSION_UP	196	2.065486	0	0.00431584
REACTOME_UNFOLDED_PROTEIN_RESPONSE	77	2.063667	0	0.00436145
CHAUHAN_RESPONSE_TO_METHOXYESTRADIOL_UP	50	2.063804	0	0.00440938
GSE22886_UNSTIM_VS_IL2_STIM_NKCELL_DN	196	2.057969	0	0.00451226
REACTOME_LATE_PHASE_OF_HIV_LIFE_CYCLE	102	2.0596	0	0.00451575
GCNP_SHH_UP_LATE.V1_UP	177	2.060253	0	0.00452608
DNA_REPLICATION	100	2.058307	0	0.0045414
GSE15215_CD2_POS_VS_NEG_PDC_DN	198	2.056371	0.001961	0.00455216

RESPONSE_TO_ENDOGENOUS_STIMULUS	195	2.046859	0	0.00456918
GSE27786_BCELL_VS_MONO_MAC_UP	194	2.048071	0	0.00457211
GSE22886_UNSTIM_VS_STIM_MEMORY_TCELL_DN	197	2.047344	0	0.0045758
GSE3982_NEUTROPHIL_VS_BCELL_DN	196	2.058475	0	0.00457977
PEART_HDAC_PROLIFERATION_CLUSTER_DN	76	2.048443	0	0.00459774
APPIERTO_RESPONSE_TO_FENRETINIDE_DN	50	2.055294	0	0.0046019
UDAYAKUMAR_MED1_TARGETS_UP	133	2.048881	0	0.0046127
SHIPP_DLBCL_CURED_VS_FATAL_DN	43	2.045286	0	0.00465365
MATTIOLI_MGUS_VS_PCL	109	2.049027	0	0.00465502
GSE7460_CTRL_VS_TGFB_TREATED_ACT_FOXP3_HET_TCONV_UP	188	2.052417	0	0.00465942
E2F1_UP.V1_UP	186	2.05129	0	0.00467773
HOLLEMAN_VINCRISTINE_RESISTANCE_ALL_DN	19	2.053067	0.003937	0.00467924
GSE17721_CPG_VS_GARDIQUIMOD_1H_BMDM_DN	197	2.049844	0	0.00468706
GSE17721_LPS_VS_PAM3CSK4_4H_BMDM_DN	195	2.049036	0	0.00468794
GSE17721_POLYIC_VS_GARDIQUIMOD_12H_BMDM_DN	188	2.053778	0	0.00468843
GSE13484_UNSTIM_VS_YF17D_VACCINE_STIM_PBMC_UP	197	2.053575	0.001988	0.00469385
SANSOM_APC_TARGETS_REQUIRE_MYC	199	2.041031	0	0.00470021
GSE1460_INTRATHYMIC_T_PROGENITOR_VS_NAIVE_CD4_TCELL_CORD_BLOOD_UP	195	2.051536	0	0.00470966
GSE15930_NAIVE_VS_48H_IN_VITRO_STIM_CD8_TCELL_DN	194	2.043831	0	0.00471713
GSE3982_NEUTROPHIL_VS_BASOPHIL_DN	198	2.041228	0	0.00473075

GSE10325_LUPUS_BCELL_VS_LUPUS_MYELOID_UP	194	2.050191	0	0.00473128
REACTOME_HIV_LIFE_CYCLE	115	2.042349	0	0.00473246
CAFFAREL_RESPONSE_TO_THC_24HR_5_DN	57	2.040378	0	0.00474929
REACTOME_HIV_INFECTION	194	2.042478	0	0.00475367
RESPONSE_TO_DNA_DAMAGE_STIMULUS	159	2.039764	0	0.0047598
GSE24634_TREG_VS_TCONV_POST_DAY3_IL4_CONVERSION_UP	198	2.041501	0	0.00476144
GSE22886_NEUTROPHIL_VS_DC_DN	199	2.039272	0	0.00478361
RRNA_METABOLIC_PROCESS	16	2.037801	0	0.00480522
GSE17721_CTRL_VS_LPS_4H_BMDM_UP	193	2.038513	0	0.00481256
GSE17721_0.5H_VS_8H_CPG_BMDM_UP	195	2.035938	0	0.00487455
GSE32423_MEMORY_VS_NAIVE_CD8_TCELL_IL7_DN	188	2.035964	0	0.00491234
LIGASE_ACTIVITY	95	2.036412	0	0.00491428
RNA_CATABOLIC_PROCESS	22	2.03599	0	0.00493109
GSE17721_PAM3CSK4_VS_GADIQUIMOD_6H_BMDM_UP	194	2.034405	0	0.00497094
GSE22886_NAIVE_CD4_TCELL_VS_12H_ACT_TH1_DN	198	2.033518	0	0.00501311
DNA_REPAIR	124	2.031624	0	0.0050236
GSE1460_DP_THYMOCYTE_VS_NAIVE_CD4_TCELL_CORD_BLOOD_UP	195	2.027974	0	0.00503725
HOLLEMAN_PREDNISOLONE_RESISTANCE_B_ALL_UP	22	2.031866	0	0.00506109
GSE17721_LPS_VS_PAM3CSK4_6H_BMDM_DN	191	2.028148	0	0.00506629
LI_WILMS_TUMOR_VS_FETAL_KIDNEY_1_DN	161	2.026351	0	0.00506839

MACROMOLECULE_CATABOLIC_PROCESS	135	2.030409	0	0.00508643
REACTOME_SIGNALING_BY_TGF_BETA_RECEPTOR_COMPLEX	60	2.028357	0	0.00508935
GSE36392_TYPE_2_MYELOID_VS_MAC_IL25_TREATED_LUNG_DN	191	2.031929	0	0.00509043
TRANSCRIPTION_INITIATION_FROM_RNA_POLYMERASE_II_PROMOTER	29	2.026359	0.003839	0.00510384
REACTOME_GLUCOSE_METABOLISM	63	2.028658	0.00202	0.00511202
GSE17974_CTRL_VS_ACT_IL4_AND_ANTI_IL12_12H_CD4_TCELL_DN	192	2.028891	0	0.00513911
NUCLEOTIDYLTRANSFERASE_ACTIVITY	47	2.029026	0	0.00517662
PID_P53REGULATIONPATHWAY	59	2.024776	0.001942	0.00523472
GSE17974_1H_VS_72H_UNTREATED_IN_VITRO_CD4_TCELL_DN	197	2.020952	0	0.00526651
GSE18791_CTRL_VS_NEWCASTLE_VIRUS_DC_14H_UP	198	2.024447	0	0.0052754
FUJII_YBX1_TARGETS_DN	198	2.02139	0	0.00527661
GSE17721_CTRL_VS_LPS_8H_BMDM_UP	193	2.022046	0	0.00530482
GROSS_HYPOXIA_VIA_ELK3_AND_HIF1A_DN	100	2.021547	0	0.00530553
GSE3982_BASOPHIL_VS_TH2_DN	195	2.022263	0	0.00531226
GSE27786_LIN_NEG_VS_CD8_TCELL_UP	191	2.019342	0	0.00533264
XU_HGF_TARGETS_INDUCED_BY_AKT1_48HR_DN	27	2.018727	0	0.00533638
TRANSLATION_INITIATION_FACTOR_ACTIVITY	24	2.022329	0	0.0053484
FLECHNER_PBL_KIDNEY_TRANSPLANT_OK_VS_DONOR_UP	149	2.017314	0	0.0054701
GSE15930_NAIVE_VS_24H_IN_VITRO_STIM_IL12_CD8_TCELL_DN	195	2.015525	0	0.00559023
MITOTIC_CELL_CYCLE	151	2.015876	0	0.00559529

GSE31082_DN_VS_DP_THYMOCYTE_UP	188	2.015132	0.001965	0.00559739
LEE_CALORIE_RESTRICTION_MUSCLE_DN	50	2.016024	0	0.00560477
PID_TELOMERASEPATHWAY	68	2.012894	0	0.0057169
GSE22886_IGM_MEMORY_BCELL_VS_BLOOD_PLASMA_CELL_DN	197	2.012698	0	0.00573127
GSE7764_IL15_TREATED_VS_CTRL_NK_CELL_24H_UP	191	2.012277	0.001965	0.00573473
GSE24634_NAIVE_CD4_TCELL_VS_DAY5_IL4_CONV_TREG_DN	199	2.013017	0	0.00573523
MICROTUBULE_ORGANIZING_CENTER_PART	19	2.011674	0	0.00574731
GSE12845_IGD_NEG_BLOOD_VS_DARKZONE_GC_TONSIL_BCELL_DN	199	2.008161	0.003922	0.00584568
GSE24634_TEFF_VS_TCONV_DAY3_IN_CULTURE_UP	195	2.008922	0	0.00587067
GSE28237_FOLLICULAR_VS_EARLY_GC_BCELL_DN	191	2.008175	0	0.00587986
REACTOME_RNA_POL_II_TRANSCRIPTION	102	2.008989	0	0.00588731
GSE29618_PRE_VS_DAY7_POST_TIV_FLU_VACCINE_PDC_DN	197	2.008358	0	0.0059079
MITOCHONDRIAL_ENVELOPE	95	2.009067	0.001961	0.00591388
GSE27786_BCELL_VS_NEUTROPHIL_UP	192	2.006471	0	0.00592356
SPLICEOSOME	50	2.009313	0	0.00593155
SCHLOSSER_MYC_AND_SERUM_RESPONSE_SYNERGY	31	2.006896	0	0.00593224
GSE5960_TH1_VS_ANERGIC_TH1_UP	194	2.009563	0	0.00594863
REACTOME_DEADENYLATION_OF_MRNA	19	2.005338	0	0.00600736
CHROMOSOME_ORGANIZATION_AND_BIOGENESIS	123	2.003089	0	0.00607011
GSE3982_NEUTROPHIL_VS_CENT_MEMORY_CD4_TCELL_DN	196	2.003532	0	0.00607614

HESS_TARGETS_OF_HOXA9_AND_MEIS1_UP	64	2.003721	0	0.00608331
GSE24634_TREG_VS_TCONV_POST_DAY10_IL4_CONVERSION_UP	197	2.004116	0	0.00609023
DACOSTA_UV_RESPONSE_VIA_ERCC3_TTD_UP	64	2.002618	0	0.00609442
REACTOME_RNA_POL_II_PRE_TRANSCRIPTION_EVENTS	60	2.000742	0	0.00618248
NUCLEAR_BODY	31	2.001085	0	0.00619478
BURTON_ADIPOGENESIS_5	121	2.00116	0.001957	0.006229