Rheumatoid arthritis: alterations in DNA methylation patterns in CD4+T-cells

Understanding pathogenesis and developing a qMSP biomarker for classification

Rujiraporn Pitaksalee

Submitted in accordance with the requirements for the degree of
PhD

The University of Leeds
School of Medicine

February, 2021

The candidate confirms that the work submitted is his/her own, except where work which has formed part of jointly-authored publications has been included. The contribution of the candidate and the other authors to this work has been explicitly indicated below. The candidate confirms that appropriate credit has been given within the thesis where reference has been made to the work of others.

Chapter 4- Genome wide methylation data analysis contains the work included in a jointly authored publications.

The publication : "Differential CpG DNA methylation in peripheral naïve CD4+ T-cells in early rheumatoid arthritis patients", R. Pitaksalee, A. N. Burska, S. Ajaib, J. Rogers, R. Parmar, K. Mydlova, X. Xie, A. Droop, J. S. Nijjar, P. Chambers, P. Emery, R. Hodgett, I. B. McInnes & F.Ponchel. Clinical Epigenetics volume 12, Article number: 54 (2020) DOI: 10.1186/s13148-020-00837-1

My contributions incudes the analysis of DNA methylation data with some support/training (AD, PC JSN), the development of a system to select differentially methylated CpGs (RH), the validation of the DM CpG by ELISA. I then collaborated with a Master student to performed the TNF bisulfite sequencing (JR), and with others to establish differentially gene expression lists from publicly available data (SA, KM), which I then compared with differentially methylated genes. The section of flow cytometry was mainly my supervisor's work (FP) with the help of a student and research technician (XX, RP). Clinical data were retrieved by AB who helped with many practical aspects. The overall supervision was provided by FP, while IBMcI contributed to the scientific discussion of this project. FP/IBMcI were CO-investigators on the grant which provided support for the acquisition of the genome wide data (EuroTEAM).

Acknowledgements

Completion of this thesis was possible with the support of many people. I would like to take this opportunity to express my gratitude to them.

First and foremost, I would like to express my sincere thanks to my supervisor, Dr Frederique Ponchel who gave me an opportunity to do this project. I would like to thank for all her help and support throughout my PhD. I appreciated her valuable advice, consistent encouragement and her understanding.

I would like to thank Dr Richard Hodgett, my co-supervisor who opened the door to data analytics for me and supervised me especially at the beginning of my study in Leeds.

I would like to thank all group members for their support and good collaboration. I has a great time working here. My grateful thanks are also extended to Leeds Institute of Rheumatic and Musculoskeletal Medicine and Welcome Trust Brenner Building's staff as well as the patients and volunteers who donate blood samples for my project. Without them, this project would have been impossible.

I would like to acknowledge the Royal Thai Government Scholarship for giving me the opportunity to do my PhD in the UK, with full support for all the cost of my study.

Finally, I would like to thank my family, who always gives me the greatest love, support and help me get through any difficult time. Also, thank to my friends both in Leeds and Thailand for their warm friendship and encouragements.

Abstract

Alterations in DNA methylation patterns have been related to several diseases, including Rheumatoid Arthritis (RA). CD4+T-cells are critical players for the early pathogenesis of RA. I hypothesise that modification of DNA methylation in CD4+T-cells happen early in RA and contribute to the disease progression by altering important physiological pathways. The aims of my thesis are

1) to gain more understanding of early events/pathways in RA pathology by studying genome-wide DNA methylation

2) to select potential CpG candidates for the development of a biomarker for the prediction of clinical outcomes.

For the first aim, Illumina methylation genome-wide array data were analysed in naïve and memory CD4+T-cell and monocytes from 6 healthy control (HC) and 10 early, drug naïve RA patients. DNA methylation pattern in naïve CD4+T-cell confirmed the involvement of several pathways (mainly IL6/STAT3 linked to TNF-α, and IFN signalling genes) in the early disease pathogenesis and importantly discovered novel pathways such as dysregulation of the commitment of Th17 polarisation in naïve cells. My findings suggested a novel disease mechanism model in which IL6 induces atypical differentiation in a small subset of naïve CD4+T-cells, potentially associated with a subset of cells previously observed *in vivo* in RA patients by my supervisor.

For the second aim, I used several publicly available methylation datasets to develop selection strategies to identify CpGs candidate to develop as a biomarker assay for RA classification using a quantitative Methylation-Specific PCR (qMSP) technique. A *TNF* qMSP assay was successfully developed. It detected difference in methylation levels between RA and other arthritis with a good classification performance (n= 284, AUROC = 0.171 (95%CI: 0.115 - 0.227). This assay also showed potential to predict response to Methotrexate (pilot study). Further validation with a larger cohort will be necessary to included such assay in the management of early RA.

Table of Contents

List of Tables

List of Figures

List of Abbreviations

Ab	Antibody
ACPA	Anti-Citrullinated Protein Antibody
Ag	Antigen
AIDS	Auto Immune Diseases
ANA	Anti-Nuclear Antibody
anti-CarP	Anti-Carbamylation
APCs	Antigen-Presenting Cells
AUROC	The area under the ROC curve
autoAbs	Auto Antibodies
ChA	Chapel Allerton
CpG	Cytosine-phosphate-Guanine
CRP	C-Reactive Protein
CTLA-4	Cytotoxic T-lymphocyte antigen 4
DAS	Disease Activity Score
DCs	Dendritic Cells
DEG	Differentially Expressed Genes
DM	Differential Methylation / Differentially methylated
DMARD	Disease-modifying anti-rheumatic drug
DMR	Differentially Methylated Region
DNMTs	DNA methyltransferases
EAC	Early Arthritis Clinic
ESR	Erythrocyte Sedimentation Rate
FACS	Fluorescence-Activated Cell Sorting
FDR	False Discovery Rate
FLSs	Fibroblast-Like Synoviocytes
FN	False negatives
FP	False positives
GEO	Gene Expression Omnibus
HC	Healthy Control
HLA	Human leukocyte antigen
IA	Inflammatory Arthritis
IACON	Inflammatory Arthritis disease CONtinuum
IFN	Type-I interferon
IFNs	Interferons
IL	Interleukin
IMID	Immune-Mediated Inflammatory Diseases
IRC	Inflammation Related Cells
mAbs	monoclonal antibodies
MDS	Multidimensional scaling
MHC	Major Histocompatibility Complex
MMPs	Matrix MetalloProteinases
MTX	methotrexate

MWU	Mann-Whitney U
NF-κB	Nuclear Factor κB
NPV	Negative predictive value
NSAIDs	Non-Steroidal Anti-Inflammatory Drugs
NTC	No Template Control
OR	Odds ratios
PB	Peripheral Blood
PBMC	Peripheral Blood Mononuclear cells
PPV	Positive Predictive Value
PTM	Post-Translational Modification
QC	Quality Control
qMSP	quantitative Methylation-Specific PCR
RA	Rheumatoid arthritis
RANKL	Receptor Activator of Nuclear factor Kappa-B Ligand
RF	Rheumatoid Factor
ROC	Receiver Operator Curve
ROS	Reactive Oxygen Species
SE	Shared Epitope
SF	Synovial Fibroblast
SJC	Swollen Joint Count
SLE	Systemic Lupus Erythematosus
SNP	Single-Nucleotide Polymorphism
TCR	T-cell Receptor
TF	Transcription Factor
Tfh	Follicular helper T cells
TGF	Transforming growth factor
Th	Helper T-cells
TIMPs	Tissue Inhibitors of MetalloProteinases
TJC	Tender Joint Count
Tm	Melting temperature
TN	True negatives
TNF	Tumor necrosis factor
TP	True positives
TREC	TCR Rearrangement Excision Circles
T-reg	Regulatory T-cell
UA	Undifferentiated Arthritis
VEGF	Vascular Endothelial Growth Factor
WB	Whole Blood
*R	Receptor
α	Alpha
β	Beta
γ	Gamma
Δ	Delta

Chapter 1 General introduction

1.1 Rheumatoid arthritis

Rheumatoid arthritis (RA) is a chronic inflammatory disease that primarily affects synovial joints. It affects approximately 0.5-1.5% of the population in the UK and 0.5-1% worldwide (1). Prevalence is three-times higher in women than in men and it increases with age. The average age of onset lies between the ages of 45 (during active life) and up to 65 years, but it can also affect younger people. Frequent signs and symptoms include pain, swelling, stiffness of joints, and fatigue (2).

RA is part of a group of immune-mediated inflammatory diseases (IMID) and considered to be an autoimmune disease. It is characterized by persistent synovial inflammation (synovitis) and swelling of multiple joints (hands, knees and feet) in a symmetrical pattern, association with a particular Major Histocompatibility Complex (MHC) (HLA-DR), autoantibody production (Rheumatoid factor (RF) and anti-citrullinated protein antibody (ACPA)), cartilage and bone destruction, as well as systemic features (e.g. cardiovascular, pulmonary, and depression) (3).

There is not yet a cure for RA. It is a life-long condition. The treatment of RA aims to relieve patients from suffering and slowing down the disease course. However, despite the improved RA management strategies available nowadays, many patients still cannot attain the full benefits of current treatment options. This life-long disease also has a considerable socio-economic burden with direct and indirect costs of over 2 billion £ /year in the UK (4). The costs are expected to increase further due to the rising need for more effective, but more expensive treatments notably biological compared to the first-line synthetic drug. The optimal management of RA as a life-long condition is a priority for patients, health authorities and societies. It relies mainly on early diagnosis and intervention, before the damage and disability become irreversible.

1.1.1 Etiology and pathophysiology of RA

The exact cause of RA remains unclear. RA has a strong genetic background with heritable rate from parents around 40-65% for APCA+, while lower in APCA- disease. However, the monozygotic twin studies show that the chance of both having RA is only 15-30%, suggesting other non-coding factors play an important role in susceptibility (5). RA is hypothesised to be the effect of genetic susceptibility together with environmental triggers (5).

There is a long known, strong association between RA and the human leukocyte antigen (HLA) also known as major histocompatibility complex (MHC)) loci, especially with the *HLA-DR* gene of the B1*01 B1*04 and B1*10 alleles (6). These alleles are coding for chains of MHC molecules that may contain a common amino acid motif, known as the shared epitope (SE); which is significantly associated with the risk of developing RA. The SE effect on RA is thought to involve several mechanisms which are still not completely understood but may ultimately promote autoreactive immune responses (3). Evidence up to now suggests that it could promote autoreactive immune responses via (i) shaping the T-cell repertoire during thymic selection (7-9), (ii) antigen presentation with alteration in peptide affinity (10, 11) or (iii) increasing T-cell senescence(12). Alternatively molecular mimicry of microbial proteins (13) as well as a potential pro-inflammatory signalling function of the SE itself are further proposed effects unrelated to the SE activity in antigen recognition (14).

With the advance of genome-wide association studies (GWAS), >100 additional RA-associated loci have been reported although with weaker associations (15-20). Most of these loci are associated with adaptive immunity and inflammatory pathways, particularly, stimulation, activation and functional differentiation of T-cells, implicating nuclear factor κB (NF-κB)–dependent signalling and other cytokine cascades (3). Many others are related to antigen presentation to T-cells.

Environmental factors know to be of high risk to RA include smoking, gender-related factor (as well as contraceptives intake), viral infection (e.g. Epstein-Barr virus, parvovirus), bacterial infection (e.g. Proteus and Mycoplasma) and exposure to certain substances (such as cigarette smoke, mineral oils) (21-27). Other factors such as gastrointestinal microbiome, or experience with adverse or traumatic life events could also influence RA (28-30). These environmental exposures are thought to trigger RA development in those individual who have a genetic susceptibility. For example, interaction between smoking and other forms of bronchial stress (e.g. exposure to silica) increase the risk of RA among individuals with susceptibility associated with the *HLA–DR* SE alleles (31).

The exact mechanism of how an interaction of environmental factors with genetic susceptibility leads to RA remains unclear but an overall key process resulting in **loss of self-tolerance** is needed. Despite a lot of research, no classic antigen (i.e. protein) has ever been clearly associated with RA although in the past 10 years, a breakthrough defined post-translational modification (PTM) of certain protein epitopes (harbouring a consensus sequence (32)) as potential triggers of the auto-antibody response. These PTM are non-specific to the disease and occur as natural responses to cellular stress and inflammation (33, 34) while the immune response to this PTM (i.e. the development of IgG autoantibodies) is RA-specific (i.e. not present in health or other diseases, despite a natural IgM repertoire of autoAb present on ~30% of people (35, 36).

The consequence of this immune response developed against a modified self-proteins includes antibody against citrullination, called anti-citrullinated protein antibodies (ACPA), carbamylation (anti-CarP), oxidation (glycation, oxidation, ..etc) (37-39)). Another type of auto-antibody against the Fc portion of IgG, known as the Rheumatoid factor (RF) is also associated with RA (40, 41).

The presence of circulating ACPAs (42), as well as other antibodies such as RF, anti-Carp and anti-oxidised collagen (38, 40, 43) and circulating pro-inflammatory cytokines and chemokines (44) can be detected up to 15 years before clinical disease onset (45), which points to immune activation during a preclinical phase to the disease, which is now a well-recognised step wise process (46, 47). However, there is still no clear conclusion about how the **loss of self-tolerance** at the **systemic level,** can lead to a localized inflammatory response in the joint. **A second series of events** needs to happen, for which the cells involved and the possible triggers are not yet known although several hypotheses have been proposed, including dysregulation of T-cell differentiation, a role for cytokines, netosis, pain, osteoclast, Interferon-signalling and/or Th17-cells (3, 47-50).

The clinical phases of RA, when signs and symptoms develop, takes place in the joints. Synovitis (i.e. inflammation of the Synovial membrane) results from lymphocytes infiltration into the synovium compartment (tissue and fluid). Both adaptive immune cells (e.g. T-cells and B-cells), innate immune cells (e.g. monocytes/macrophages, mast cells, neutrophil and others), as well as resident cells (e.g. synovial fibroblasts (SF), chondrocytes, and osteoclasts) contribute to a cell-cell interaction network that promotes an inflammatory cascade, tissue remodelling and tissue damage. The release of chemokine recruits more immune cells. Cytokines induce the activation of macrophage-like synoviocytes and the activation/proliferation of synovial cells, (notably fibroblast-like synoviocytes

(FLSs)) causing an expansion of the synovial membrane and the release of pro-inflammatory cytokines, and other substances (such as matrix metalloproteinases (MMPs), tissue inhibitors of metalloproteinases (TIMPs), protease and nuclear factor κB ligand (RANKL)) promoting bone and cartilage damage. FLS (as well as immune cells) also have the potential to migrate from joint to joint to propagate disease (51). Furthermore, certain cytokines and in particular RANKL promote bone erosion by activation of osteoblast and chondrocytes (52). Other events during the disease process are micro-environmental changes, such as the expression of adhesion molecules in synovial micro-vessels, the induction of angiogenesis (via secretion of vascular growth factors) in the synovial membrane allowing for even more immune cells infiltration (53). The prolonged inflammation promoting cartilage damage and bone erosion leads to the deformity and malfunction of the joints and other systemic consequences (3, 54-56).

A summary of the development and progression of RA from the pre-clinical phase to clinical symptoms is shown in Figure 1-1.

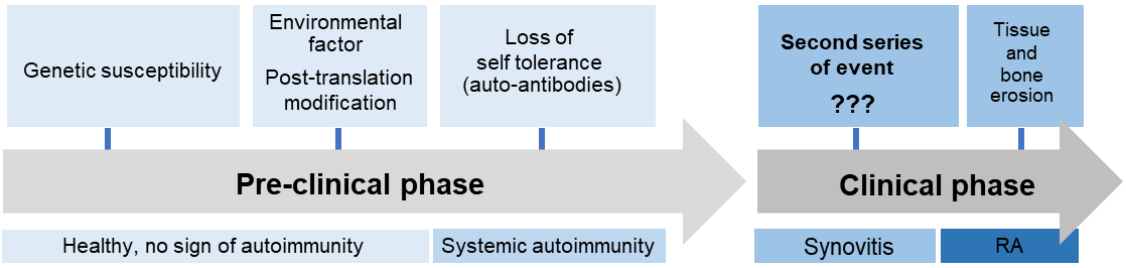


Figure 1-1 Development and progression of RA from a healthy individual to a RA diagnosis. In the preclinical phase, genetic susceptibility, and environmental factors could both contribute to the risk of developing RA. The post translational modification of self-protein by exposure to a high-risk environment is thought to promote the loss of self-tolerance, especially in the individual with genetic susceptibility. This induced an auto-immune response as detected by the appearance of auto- autoimmunity against modified self-proteins such as ACPA, anti-Carp and anti-oxydised collagen. This can occur years before onset the clinical symptoms (systemic autoimmunity). The transition from pre-clinical phase to clinical phase which suggests the localisation of the immune reaction to joints occurs after a second series of events, which remains unclear. This include a phase of arthralgia (joint pain with no clinical evidence of inflammation) still considered pre-clinical and then the development of measurable symptoms notably inflammatory acute phase markers (CRP, ESR) and swelling of the joint (not only pain). In the clinical phase, there is the infiltration of immune cells in joints and initiation of the inflammatory cascade. Individuals whose symptoms persist for 6 weeks and meet the RA classification criteria will be considered early RA. The disease still continues to progress if uncontrolled by the initiation of a 1st line treatment (towards established RA), leading to permanent tissue damage and bone/cartilage erosion.

1.1.2 Immune Cells with pivotal roles in Rheumatoid Arthritis

RA is a complex disease that involves many immune cells working and interacting together with stromal cells (i.e. SFs) in the joints. The major immune cells that are known to be involved in the perpetuation of RA pathogenesis are T-cells, B-cells and macrophages. Although they all play a significant part in disease progression, their role may to be determinant, dominant or an accessory at different stages of the disease course. My PhD project focusses on T-cells and I will therefore describe these more below.

1.1.2.1 Macrophages

Macrophages are infiltrating the RA synovial membrane. They produce a variety of pro-inflammatory cytokines and chemokines (such as TNF-α and IL1-β) (57). These pro-inflammatory factors activate a wide range of immune and non-immune cells (stromal and endothelia cells). Macrophages also release matrix-degrading enzymes and reactive oxygen intermediates that contribute to tissue damage and joint destruction (57). They can stimulate T-cell responses by acting as professional antigen-presenting cells (APCs), although in RA this is thought to be mainly the work of B-cells (58). The most abundantly produced cytokines in the RA synovium are produced by macrophages (including TNF, IL1, IL6, IL10, IL12, IL18, IL15, IL10, GM-CSF, M-CSF and TGFβ) which remain up-regulated for prolonged periods and account for the development of chronicity and the persistence of inflammation (57, 59). This stresses the important role of macrophages in the pathophysiology of RA, especially in the clinical stage when the inflammatory cascade is the main disease feature. As such, studies of RA patients synovial tissue while in clinical remission have shown that macrophages frequency and activation status are largely reduced (60). Histology/histochemistry studies notably showed that a score (OMERACT score) based on expression of macrophage markers can provide a measure of disease activity in RA (61).

1.1.2.2 B-cells

B lymphocytes play several important roles in the pathogenesis of RA including autoantibody (autoAb) production, T-cell activation via their APC function, cytokine production and involvement in bone homeostasis (62, 63).

The obvious evidence of a B-cell contribution to RA is that they are responsible for the production of APCA and other autoAbs (as discussed above, which are known to form after the triggering of PTM associated with the loss of self-tolerance). These autoAbs also contribute to immune complex formation and complement activation in the joints (64). Animal models have suggested an arthrogenic role for these autoAbs, although it is not totally clear in humans (65).

The therapeutic benefit of B-cells depletion in RA suggested that their function in RA pathogenesis involved additional mechanisms beyond autoAb production, notably as antibody levels are not changing significantly after B-cell depletion or according to response (66, 67). B-cells can promote T-cells activation through their efficient capacity for antigen-presenting and the expression of costimulatory molecules. Growing evidences show that B cells also contribute to the joint expression of several chemokines and cytokines (such as IL12, IL23, as well as IL1-β and TNF-α) that promote lymphocytes activation in the joints, the formation of ectopic lymphoid structures, angiogenesis, and synovial hyperplasia (68, 69). B-cells were also reported to be a source of RANKL involved in bone homeostasis (70), while autoAb recognizing citrullinated vimentin were reported to promote the differentiation of mononuclear cells into osteoclasts (50). These showed that B-cells may have a plethora of roles in both the pre-clinical phase (with autoAb production) and in the clinical phase of RA.

1.1.2.3 T-cells

T lymphocytes play significant roles in the RA pathogenesis both in the preclinical and in the clinical phase. T lymphocytes are cells of the adaptive immune system. T-cells mature in the thymus where they are “trained” to become a functional T-cells through various processes including (i) the rearranging of the T-cell receptor (TCR) gene fragment to generate a wide variety of TCR allowing for responses to a large diversity of pathogens, (ii) a positive selection to ensure the optimal binding to MHC class I and Class II molecule to the TCR and the commitment of cells to present either a CD4+ or CD8+ costimulatory molecule, and (iii) the negative selection to ensure the tolerance to self-antigen. The mature naive T-cells leave the thymus and then circulate in the blood to the secondary lymphoid organ.

In a pathogen invasion situation, T-cells are activated by the engagement of the TCR and antigen-MHC-I or II complexes on the surface of the APC which have picked-up this pathogen, as well as express co-stimulatory signals. They then proliferate and differentiate into effector cells and maintain a small proportion of cells differentiated into memory T-cell.

CD4+T-cell or T-helper cells engage with MHC II on APC. Once activated they can differentiate into different subtypes of effector T-helper cells based on the context in which APC were activated (i.e., B-cells or Macrophages and the type of pathogen), whereas CD8+T-cell, or T-killer cells which engage with MHC-I are mainly responsible for the elimination of pathogens that have infected cells.

CD4+T-cells are intimately involved in cell-mediated immune responses and have long been known to be involved in RA development and progression. In 1982, the “T-centric” hypothesis in which T-cells get activated by a hypothetical arthrogenic agent (viral, bacterial or autoantigen) and then orchestrate the inflammatory response and development of RA (71) was raised. Later in the 1990’s, a “cell-network” hypothesis involving macrophages (56, 72) gained more favour, and the T-cells centric hypothesis dropped out of fashion notably as no common T-cell antigen could be identified (73). More recently, advances in genomics technologies brought novel evidence supporting the strong association of T-cells with RA.

Genetic evidence

Strong evidence for the role of T-cells in RA pathogenesis is that many RA susceptibility loci are T-cell related genes, either directly engaged in the T-cells activation process or in related signalling pathways. The *HLA-DRB* gene, which

codes for the SE has the strongest association with RA (6). This SE is part of the MHC-class-II molecule expressed on APC, which directly contacts the TCR complex on CD4+T-cell during their maturation (positive and negative selection in the thymus) as well as during the antigen activation process in mature T-cells in the periphery. How this SE affects T-cells is still not clear but some evidence shows that it could shape the TCR repertoire and possibly promote autoreactive immune responses as discussed above (3, 7-14, 74).

A second major breakthrough from these genetic studies was to bring more insight into RA aetiology. In the past decade, GWASs in various RA patients populations and their meta-analysis together with post-GWAS multi-Omics data uncovered over 100 RA-susceptibility loci (16-19). Molecular pathway analysis of these 100 non-MHC RA-risk loci revealed that they are lying mostly within genes involved in T-cell biology related pathways, also in antigen presentation pathways and cytokine signalling pathways (16, 19, 75-78). The most well-studied RA susceptibility loci/genes are, for example:

- *PTPN22* (Protein tyrosine phosphatase, non-receptor type 22) which encodes an enzyme involved in signalling pathways associated with the immune response, alters the responsiveness of T and B cell receptors, associated with several autoimmune diseases.
- *CTLA4* (cytotoxic T-lymphocyte-associated protein 4 or CD152) which encodes a receptor that functions as an immune checkpoint and downregulates immune responses on T cells after activation.
- *IL2RA* and *IL-2RB* (Interleukin-2 receptor alpha and beta chains) which involve in stimulating T-cell proliferation.
- *STAT4* (Signal transducer and activator of transcription 4) which required for the development of Th1 cells from naïve CD4+ T cells and IFN-γ production in response to IL-12.
- *PADI4* (Peptidyl Arginine Deiminase 4) coding for the enzyme responsible for citrullination, involved in the cellular response to stress, creating potential antigens that bind to RA-associated MHC proteins.

Recently (in 2020), the largest so far Trans-ethnic GWAS meta-analysis which recruited more than 300,000 individuals (RA and controls) from European and Asian ancestors has been performed (79). The study identified 11 new loci and confirmed 71 known- non-HLA susceptibility loci, with another 90 association signals related to RA. The majority of variants in RA were shared between the two ancestries groups, which have highly distinct linkage disequilibrium architecture, suggesting a true insight into RA aetiology. They also integrated the

GWAS results with other omics data for transcription factor binding sites and histone modification marks, performing enrichment analysis of RA variants. Large heritability variants presented preferential relationships with binding sites of the transcription factors associated with T-cell receptor signalling. The histone modification marks (such as H3Kme3 involved in chromatin accessibility), were strongly associated with RA-risk loci specific of CD4+T-cells including stimulation, memory and/or regulatory T-cells in agreement with previous trans-ethic GWAS meta-analyses performed on >100,00 individuals reported earlier (19). This further emphasized the key role for T-cells in the initiation and/or perpetuation of RA.

Note that although more than 100 RA-risk loci have been identified, a large proportion of RA heritability is still unexplainable (19, 78, 79). The Discovery of new RA susceptibility loci and the incorporation of other biologically relevant analyses such as transcriptomic and epigenetic analyses remains necessary for a more complete understanding of RA pathogenesis (80).

Evidence of abnormalities in the CD4 T cell repertoire

Contraction in the TCR diversity of CD4+T-cells and clonal populations in the CD4 compartment were reported in RA (81).

Clonal expansion was, at first explained as a consequence of specific responses to synovial self-antigens, however this hypothesis of an antigen drive autoimmunity in RA is unlikely to be absolutely true because such clonal expansions were not limited to the memory compartment, but involves also naïve T cells (82, 83). The loss of TCR diversity and the clonal expansion appeared to be due to accelerated ageing of the immune system in RA patients, whereby thymic function is lost earlier than in healthy people as suggested by the decline of TCR rearrangement excision circles (TREC, an indicator of newly generated T cells from thymus) notably in early, drug naïve RA patients (82, 84). A further reduction in diversity in the T-cells repertoire may result from a clonal expansion in order to fill the void left by CD4+ T-cell lymphopenia in RA (85). However, the autoreactivity may result from other processes as well. The abnormal CD4+T-cells repertoire was also shown to results from the shaping of TCR by the SE (possibly by presenting self-peptides to CD4+ T cells in the thymus (7, 9, 82, 86) as well as preferential binding of citrullinated peptide by the SE alleles (73). This might result in selecting TCRs that are more reactive to self-antigens and lead to autoimmunity later.

Evidence of accelerated in T-cell senescence

Premature senescence of T-cells in RA is associated with telomeres loss (12). The insufficient upregulation of telomerase activity was also reported in RA and telomere erosion was found in both the naïve and memory T-cell compartments as well as in progenitor cells (87). Loss of telomere length eventually leads to loss of gene on the end of chromosomes which can further affect the function of the cells.

Evidence of defect in T-cell differentiation

Defects of T-cell differentiation were observed in RA. The clonal expansion of unusual CD4+T-cells (CD4+ CD28- T cells) characterized by lacking CD28 cell surface molecule, functionally important for T-cells activation, have been in observed in the blood of RA patient (88). The presence of large numbers of such unusual T cells is likely to influence immune responsiveness and alter mechanisms of inflammation (88-90).

Defects in CD4+T-cells differentiation was also studied by my supervisor research group. The loss of naïve cells and the appearance of atypical CD4+T-cells that are expressing both naïve and memory differentiation markers (CD45RB, CD45RA, CD45RO and CD62L), namely inflammation related cells (IRC) (91-93), was observed in the blood of early, drug naïve as well as established RA patients. The correlation between the frequency of this unusual cell subsets and levels of C-reactive protein (CRP) and the reduced TREC content compared to naive cells, while still being high in IRC, suggested that inflammation drives the proliferation of naive T cells and encourages their differentiation into atypical IRC. These are also hyper-responsive to antigen and mitogen stimulation while remaining naïve with respect to an antigen experience (92, 93). They also persist in remission (94, 95).

Evidence of abnormal T-cell polarization

Naïve CD4+T-cells differentiate and polarize to different effector; helper T-cells subsets (Th, e.g. Th1, Th2, Th17, Th9, Th22), follicular helper T-cells (Tfh) or inducible Regulatory T-cells - (iT-reg), according to the stimuli they receive. Each effector subset release a different pattern of cytokines Figure 1-2. The abnormal polarization of these cells or the cytokines/chemokines releasing from them are important in cell-mediated immunity and can contribute to the pathogenesis of RA.

Th1 cells release pro-inflammatory cytokines such as IFN-γ, IL2 and TNF-α and promote the activation of other immune cells such as macrophages, B cells, and CD8+ cytotoxic T-cells, required for host defense against intracellular viral and bacterial pathogens. Although the pro-inflammatory cytokine of Th1 cells could contribute to inflammatory cascade in autoimmune reactions, they might not be involved in RA pathogenesis as the differentiation of naïve CD4+Tcells toward Th1 was found to be impaired in RA (96, 97) with impaired commitment and expression of T-bet, the master regulator of Th1 polarisation (98, 99).

Th2 cells produce IL4, IL5, IL9, IL13, and IL17E/IL25 which are important for the induction and development of humoral immunity and play an important role in coordinating the immune response to large extracellular pathogens. Differentiation of naive CD4+T cells towards Th2 appears intact in RA (99, 100).

Apart from the classic Th1 and Th2 subsets, newly identified cell subset such as Th17, Th9 and Th22 have also been implicated. Th17 cells have significant roles in RA pathogenesis. This subset produces potent pro-inflammatory cytokine such as IL17A, 17F, IL21, IL22 and TNF-α (101). The presence of numerous Th17-cells and high levels of IL17 in the joints of RA patients was reported (102-104). These promote the activation of fibroblasts and chondrocytes, induces expression of the osteoclast differentiating factor RANKL. Besides, Th17 are able to suppress the differentiation of iTreg which have a role in regulating or suppressing other immune cells, thus shifting T-cell homeostasis toward inflammation (105). Despite this, studies about the role of Th17-cells in RA pathogenesis have suggested that Th17 cells may play a central role in the early disease-onset stages of RA from the at-risk stages up to diagnosis, but have limited effect later in the disease course (102, 106-108), notably in line with the lack of therapeutic efficacy of anti-IL17 Ab in RA (109).

An increased frequency of Th9 cells and of the expression of IL9 was reported in the synovial tissues and fluid of RA patients (110). Th9 cells released IL9 which is involved in prolonging the survival of neutrophils in the synovium, stimulated the production of MMP-9 and facilitated Th17 cell differentiation (111).

Th22 cells, which secrete IL22, are more likely to be recruited in the skin and contribute to host defines against microbial pathogens and promote tissue repair or remodeling (112).

Treg cells are responsible for immune suppression, maintaining immune homeostasis and self-tolerance, which is mediated by regulating the activity of the effector T-cells via releasing of the suppressor cytokines IL10 and TGF-β. In early and preclinical RA, Treg cell function is impaired, while Treg cell counts are

also reduced (48, 91, 113, 114). The balance between Th17/Treg is tilted toward more Th17 cells, suggesting an association with RA development (115).

Tfh cells release cytokine such as IL4, IL21, and IFN-α, promote the survival and proliferation of B cells in germinal center and their production of antibodies. An increased frequency of peripheral Tfh cells was reported in RA patients (116). This high level of Tfh also correlated with increasing ACPA levels suggesting the possible involvement of Tfh cells in the disease progression of RA (116).

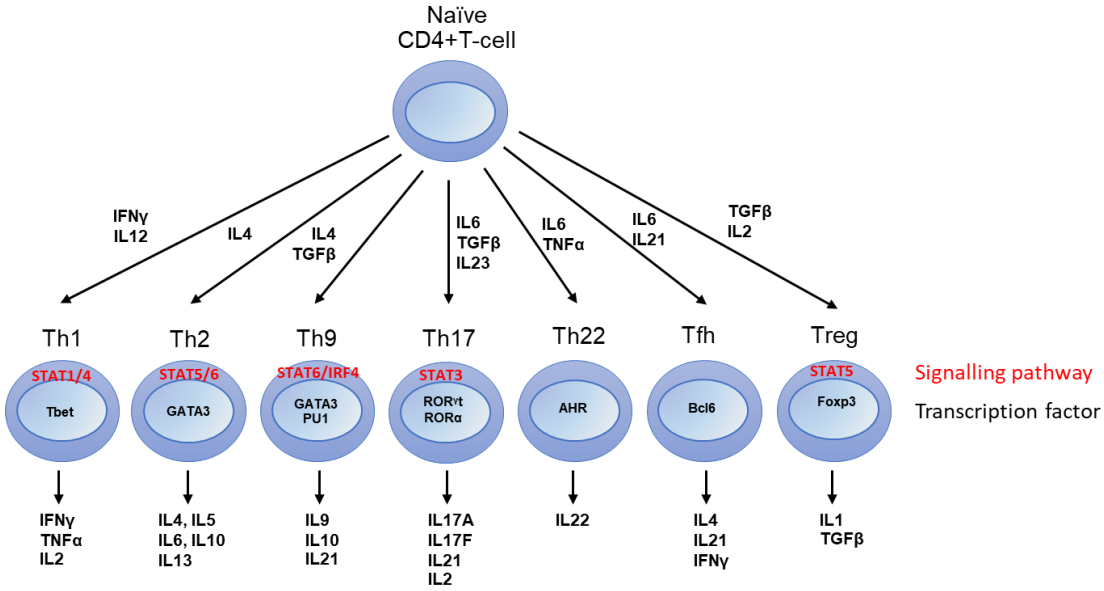


Figure 1-2 CD4+T-cells subsets Naïve CD4+T-cells can differentiate into a number of specific cells subset according to the signals provided by the microenvironment (i.e. polarising cytokine milieu). The differentiation/polarisation requires signalling through cytokine receptor/JAK/STAT complex resulting in the induction of the expression and epigenetic commitment of transcription factors as master regulator of the engagement in polarisation (Tbet, GATA3, RORyt and RORa, AHR, BCL6, Foxp3). The diagram shows required cytokine signal, transcription factors of specific cell subset included Th1, Th2, Th9, Th17, Th22, Tfh, and Treg as well as the cytokine production profile.

Evidence of disturbed cytokine production

Cytokines are responsible for the communication between the parts of the immunes system and have an important role in RA pathogenesis. Cytokines can affect proliferation, differentiation/polarisation, and the survival of T-cells, all functionalities that contribute to RA pathogenesis.

Cytokines are very important mediators in RA as proven by the success of therapies blocking their effect such as anti-TNF (117, 118), anti-IL6 (119) as well as JAK-inhibitor (120) more recently, while anti-IL1 was less effective (121, 122) and surprisingly anti-IL17 was not at all (123, 124).

Effector CD4+T-cells release cytokines that promote inflammation (e.g., TNF-α, IFN-γ) or cytokine that promotes B-cell maturation (e.g., IL6 notably). However, the main source of cytokines expressed in the synovium in established RA (e.g. TNF-α, IL6, IL1, IFN-γ) are macrophages and FLS (125). This was the main argument to suggest a passive role for T-cell in promoting inflammation (in the late 90's) however, this was observed once the disease is already fully established.

T-cells also contribute to RA pathogenesis through other cytokine production. A clear example is the production of potent pro-inflammatory cytokines such as IL17A (126), IL17F (127), IL21 (128), and IL22 (129) and TNF-α (101, 130) by Th17 cells. IL17 promotes recruitment of other cells such as monocytes and neutrophils in the synovium, which in turn can further fuel inflammation in RA (131). They also promote activation of fibroblasts and chondrocytes, and osteoclast differentiation via RANKL (132).

Type-I interferon (IFN) signalling related gene expression has been described in many Autoimmune disease (AIDs) including RA (133). There is a substantial body of evidence to indicate the contribution of type-I IFN activity in RA, however with a role for IFN-β, whilst most other AIDs place a greater emphasis on IFN-α (134-136). IFN-β in the joint is notably capable of protecting T-cells from undergoing apoptosis and thus has been associated with the development and maintenance of chronic inflammation (137). IL7 which is highly expressed in the joints of RA patients, is also a major cytokine involved in CD4+T-cell survival, and is also highly expressed in RA synovial tissue (138-140).

Evidence of alteration in cross-talk between different cell types

The most recent models of RA pathogenesis suggests that T-cell mediates/ orchestrate the action of other cells contributing to disease. T-cells have interactions with other cell types that are abundantly found in RA synovium; macrophages, FLSs as well as cells such as Dendritic cells (DCs), endothelial cells, and B-cells have been described which I summarised in Figure 1-3.

Macrophages have been thought to be central to RA due to the fact they produce much of the cytokines involved in the inflammatory cascade. More recent data demonstrated that they actually share this role with FLS which are also central to RA. They notably undergo hyperplasia and express altered levels of cytokines, chemokines driving influx of immune cells, and matrix-degrading enzymes contributing to joint destruction (2).

DCs are potential key regulators of the induction of immunity in their capacity as professional APC (141). In RA synovium, where immune cells are accumulated, T-cells can be activated through the presentation of auto-antigens by APCs, co-stimulation (e.g. via pathways dependent on CD28, CD154 or CD47), or the local release of cytokines (e.g. IL12, IL6, IL21) by adjacent cells (142). In turn, T-cells can themselves produce as well as promote cytokine production by other cells, in particularly TNF-α production by macrophages as well as other cytokines such as IL1-β, IL12, IL15 IL18 through cell-cell contact interactions (83, 142, 143). Activated T-cells can induce monocyte differentiation into osteoclasts via up-regulation of receptor activator of RANKL contributing to bone erosion (144). Activated T-cells can also induce the release of TNF-α, IL15, MMPs by FLS as well as FLS proliferation (142, 145). Since the cytokines in the synovium can also induce more T-cells activation, positive feedback loops are also created which maintain a vicious cycle and perpetuate the inflammatory state.

T-cells also crosstalk to endothelial cells lining blood vessel. Endothelial cells response to Vascular endothelial growth factor (VEGF) the factor that leads to angiogenesis notably produced by cells when hypoxia develops when the synovium increase in cell content. Under the influence of IL2, Ag or hypoxia, T-cell can release VEGF, activate endothelial cells and promote angiogenesis as well as increasing expression of adhesion molecule and changing the permeability of vessels, allowing more immune cells infiltration into the synovium. VEGF in turn, also affects T-cells differentiation toward Th1 (146).

The accumulation of T cells and B cells in the synovium results in tissue organisation into lymphocyte aggregates, some with features of ectopic germinal centres (147). These can promote further recruitment of B cells via IL-21

production and promotes B-cells survival, proliferation and maturation (147), although this was demonstrated in tissue from the established disease.

All types of interactions have therefore been described in RA but whether all mechanisms are relevant during the initiation, the progression to chronicity or the established phase of disease phase remains to be determined. It is likely that a specific time line of events is directing early progression while, after the establishment of chronicity, many more pathways may get activated untimely driving the development of heterogeneity in the disease and leading to many different patient phenotypes associated with different ability to respond to certain treatments better than others.

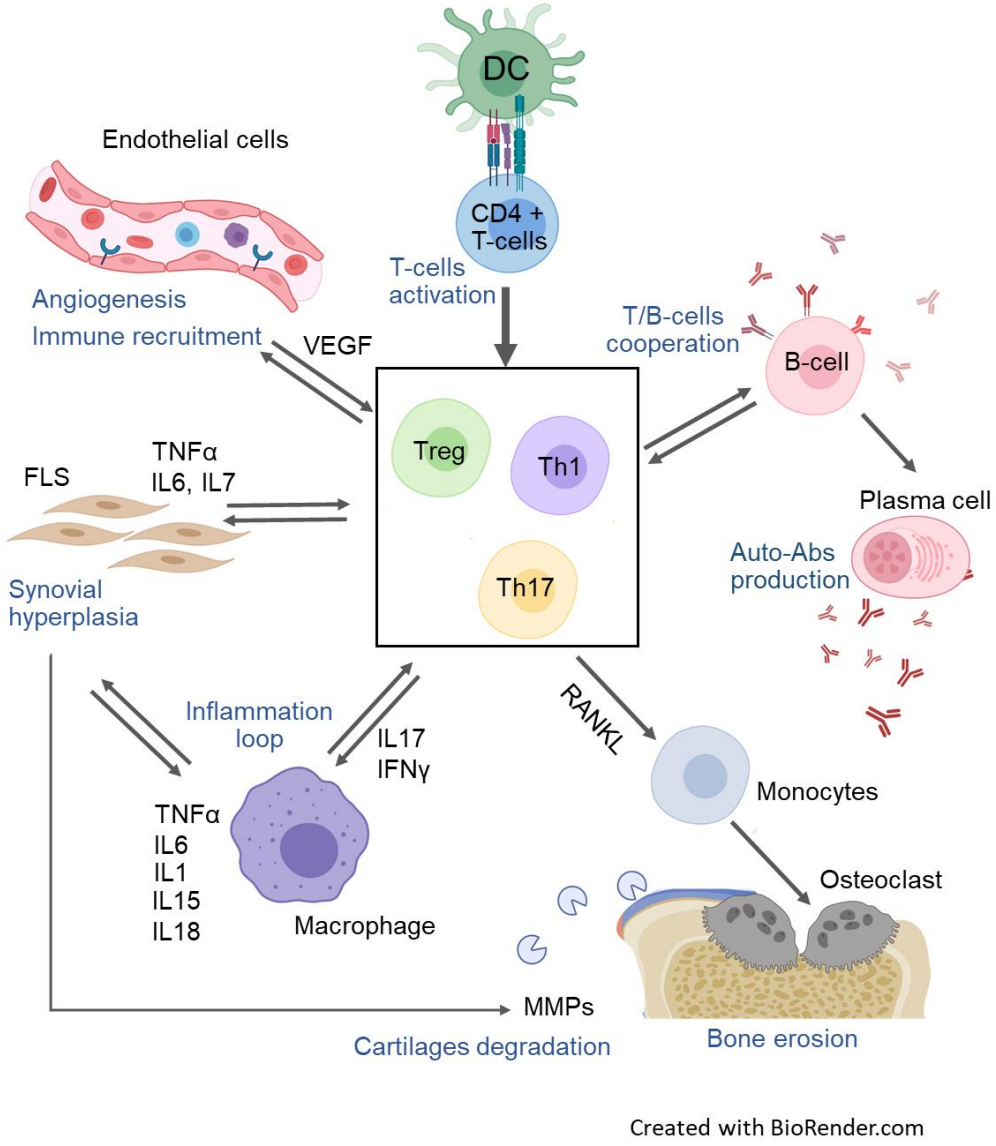


Figure 1-3 The cross-talk between CD4+T-cells and different cell types in the RA synovium.

Modern T-cell targeted therapy

Treatment for RA have evolved considerably over the past 2 decades. Historically, a number of lymphocytotoxic therapies have been employed in RA, ranging from total lymphoid irradiation to using monoclonal antibodies (mAbs) such as Campath (a T-cell depleting antibody) (148). In the early 2000's, autologous stem cell transplantation was used, in which patients received an infusion of their own pluripotent stem cells following aggressive chemotherapy, usually with cyclophosphamide (149). An initially unexpected 'side effect' of these depleting therapies was prolonged peripheral blood lymphopenia, particularly of the CD4+ T-cells subset (up to 7 years after therapy) while CD8+ T-cells and B-cells were rapidly reconstituted (150). Furthermore, lymphocytotoxic therapy was followed by oligoclonal lymphocyte expansions in the blood and synovium of RA patients, presumably from peripheral memory T-cell populations (151).

Immunosuppressive drugs have been used since the early 2000's, mainly methotrexate (MTX) (152, 153) which was then used as the standard of care for the treatment in RA a few years later (154, 155), while hydroxichlorochine (HCQ) or sulfazalasine (SSZ) were used if MTX was not well tolerated. These however mainly worked in early stage disease and allowed about half of the patients to achieve a state a clinical remission (156) that may be relatively long lasting for some, particularly if introduced very early in the disease course (i.e. less than 6 month after the development of symptoms) (157). For the other half, patients develop resistance to these agents, requiring additional drug options. With early treatment being accepted as best practice, remission was observed in many more patients and allowed the development of a new concept, the window of opportunity in early RA (158). The goal of therapy therefore became the induction of remission as early as possible in the disease course (159), with national/international guidelines (160, 161) promoting this treat to target approach (i,.e. aiming for remission) ,. This is now achieved by the introducing MTX 1st, then escalating to MTX+HCQ if remission is not achieved 3 months later and then up to triple therapy adding SSZ later (160), as the new 2018 NICE national guidelines

Finally, modern drugs were developed, some notably targeting T-cells with renewed success. The therapeutic modulation of T-cell co-stimulation (e.g.

abatacept- a CTLA4-lg fusion that block the CD28:CD80/86 costimulatory pathway(162)) or the profound immunodepletion of CD4+T-cells (e.g. keliximab, anti-CD4 monoclonal antibody (163) also confirm the significant role of T-cells in RA pathogenesis (164) while anti TNF agents (receptor blocking Ab, infliximal or adalunimab) or soluble receptor decoy (etanercpt) have now proven their efficacy for more than 10 years. The latest Ab therapy was an anti-IL6 agents (e.g. Tocilizumab, sarilumab) while the downstream signalling inhibition of the many cytokine pathways, the JAK-inhibitors (e.g baricitinib, tofacitinib) are small molecules that also both showed effective results for the treatment of patients with RA (165, 166).

In conclusion, the evidence (added to the success of drug targeting T-cells) suggested many significant roles of T-cell in RA pathogenesis. The strong link between the immunogenetic risk associated with RA, the contraction of TCR repertoire which can contribute to the loss of self-tolerance suggested a most significant role for T-cells in the initiation of disease, the pre-clinical and the early phases of RA pathogenesis. Defective of T-cells differentiation as shown in atypical CD4+T-cell subsets and the polarization of CD4+T-cells toward Th17 are then hypothesized to be part of a second hit, involved the transition from pre-clinical phase to clinical phase. Once chronicity has established itself (established RA), T-cells appear to have a more passive role as the cytokine/chemokine present in the join are mainly the production of macrophages and stromal cells.

Altogether, T-cells, B-cells, dendritic cells, endothelial cells fibroblasts and macrophages all have a role to play in RA pathogenesis, while some may be dominant at different time in the course of the disease.

Recently, the advances of single-cell technology helped pinpoint further the importance of sub-population of cells with an inflammatory state contributing to RA pathogenesis (167-169). The integration of single-cell RNA sequencing, single-cell mass-cytometry, bulk RNA sequencing and flow cytometry data for CD4+T-cell, CD8+T-cell, B-cell, monocytes, and synovial fibroblasts from tissue biopsies of RA patient (167) identified several cell sub-populations of immune cells. The new cell subsets that were observed in RA were peripheral helper T (T_{PH} PDCD1^{+}) cells and follicular helper T (T_{FH}) cells, CD8+T-cells expressing highly IFN-γ, senescent memory CD4+T-cells, autoimmune B cells

(ITGAX^{+}TBX21^{+}), pro-inflammatory monocytes and sub-lining fibroblasts (THY1(CD90)$^{+}$ HLA-DRA^{hi}). Further change in the detailed phenotypeof these sub-populations of cells seem to be associated with RA pathogenesis, for example, change of homing capacities and effector functions of CD4+T-cells, a more inflammatory status of monocytes (expression of IL1β) and fibroblasts producing IL6. More functional studies will have to be conducted to fully understand the role of these cell subsets in RA.

Nonetheless, considering the strong evidence supporting a possible role for T-cells as central to early pathogenesis in RA, this will be the focus of my PhD.

1.1.3 Unmet needs in RA

The cause and early pathogenesis of RA still remains unclear. The obvious events after the onset of disease, are an inflammatory cascade of immune cells activation, fuelling positive feedback loops. The prolonged inflammation in joints leads to the destruction of tissue, cartilage and bone resulting in irreversible structural disabilities with time and pain. Despite recent advances, there is still no cure for RA. Treatments currently still aimed to relieve inflammation (i.e. treating symptoms but not the cause) and to slow down disease progression. Taking into account the progressive stages of the disease development while the disability are irreversible and RA a life-long condition, early diagnosis and early and effective treatment to stop the progression of the disease at its very beginning are crucial to make a real difference for patents.

In clinical practice, the main useful biomarkers commonly used to diagnose RA nowadays is the presence of autoantibodies such as RF, but more specifically antibodies to citrullinated peptides (ACPA) (170). However, ACPA positivity still has limitations and around ~40% of patients at presentation (i.e. not yet meeting the diagnostic criteria for RA), do not show the presence of ACPA (171). The delays in being able to be identified patients presenting with joint inflammation symptoms as an RA patient (i.e. meeting the criteria for such classification), prevents them from receiving a suitable treatment early, with a great reduction in clinical benefit for the rest of their life. Therefore, the development of novel biomarkers that could help in diagnosing RA (either used alone or together with other markers) remains very important, particularly for a large proportion of ACPA- RA patients.

In addition, major heterogeneity exists between patients and an effective treatment for one patient may not be effective for another. This emphasizes another urgent clinical need for the discovery, validation, and development of stratification biomarker to classify patient subgroups for the better use of therapies. Elucidation the genes/ pathway involve at the beginning of RA will also help point to target for biomarker development (diagnosis/or others) and also provide more understanding of disease pathogenesis, which will be beneficial for new target for treatment.

RA was postulated to develop on a high-risk genetic background in combination with environmental exposures, notably smoking (172). Environmental risks are often causing a modification of the epigenome of cells (173). After cancer, Immune-mediated inflammatory disorders (IMIDs) and in particular, RA, became the topic of many research studies analysing epigenetic modifications. The recent analysis of the DNA methylation patterns in synoviocytes from RA patients

demonstrated a significant shift towards hypomethylation (174-176). Some of these modifications were further observed in circulating T-cells(173) suggesting that T-cells may also be the target of epigenetic modifications that may lead to disease and also have biomarker value (due to their accessibility as blood samples) while providing new understanding of the disease pathogenesis.

1.2 Epigenetic (DNA methylation)

Epigenetics regroups a number of mechanisms that regulated gene activity without a change in the DNA sequence itself. Epigenetic marks can be inherited in daughter cells (i.e. after cell division) or via the germline to the offspring. It determines which genes in a genome are turned ON or OFF thus it is very important to normal cellular processes (177, 178). A few examples are the role of epigenetic in cell differentiation (179, 180), the formation of tissue-specific cellular phenotypes (181, 182), X chromosome inactivation in female (183, 184). Although in general the epigenetic marks are stable and propagated over multiple cell divisions, at some specific DNA sites/regions they can be modified under the influence of the environment (e.g. nutrition, stress, medication), thus serving as an important mechanism for the adaptation of cells to changes in their environment (185). However, such alteration to the original patterns may also lead to pathogenic situations; especially in conditions that could not be fully explained by genetic variations (e.g. moderate concordance between identical twin studies), or conditions with a strong impact of the environment on disease development (186). The environment can indeed influence epigenetic marks over time in particular when associated with a genetic background including variants in the epigenetic machinery enzymes, which might contribute to the disease risk. This has been largely documented, particularly in cancer (179, 187-189) and more recently in IMIDs and in RA (173).

1.2.1 Epigenetic mechanism

Epigenetic mechanisms affect gene transcription without changing the DNA sequence itself (177). There are 2 major epigenetic mechanisms; DNA methylation, and histone modifications (while interference with small RNA (e.g. small interfering RNA, microRNAs, non-coding RNAs) are also sometimes considered under the epigenetic umbrella) both causing gene expression changes.

In Eukaryotic cell, Chromosomal DNA is well-wrapped around proteins, the histones, which packed into a set of eight histone proteins formed a nucleosome structure that coiled, results into chromatin to accommodate a large amount of DNA in a small space and protect the genetic material (Figure 1-4). The chromatin structure can regulate gene transcription by altering the availability of regulatory regions (e.g. promoters and enhancers) for the transcription machinery to bind to. An open chromatin structure, called **Euchromatin**, allows DNA-binding

proteins and Transcription factor (TF) to interact with regulatory DNA sequences and activate gene transcription. On the other hand, a closed condensed chromatin structure, the **Heterochromatin**, whereby the DNA is packed tightly around protein complexes, prevents TF interaction with regulatory sequences, leading to gene silencing (190-192).

DNA methylation and Histone modifications regulate gene expression via different molecular mechanisms but both aim to control the close or open chromatin structure allowing the transcription machinery to do its job, whereas miRNA regulate gene expression by either repressing the translation of already transcribed mRNA or causing the degradation of multiple target mRNAs (193, 194)

Among these mechanisms, due to its relative stability in somatic cells and the ease of experimental detection, DNA methylation is the most studied epigenetic modification at the genome-wide level both in health and in diseases (195) and is the focus of my thesis.

DNA methylation

DNA methylation is a modification of the DNA by which a methyl group is added to the 5' position of cytosine (C) forming 5-methylcytosine (5-mC). DNA methylation mainly occurs in CpG dinucleotide (Cytosine-phosphate-Guanine, CpG site, Figure 1-4). DNA methyltransferases (DNMTs) are a family of enzymes that catalyse methylation by transferring the methyl group from S-adenyl methionine (SAM) to 5-mC (Figure 1-4). DNMT3a and DNMT3b add methylation pattern to unmodified DNA (de novo DNA methylation)(196) whereas DNMT1 is primarily responsible for the maintenance of existing methylation on CpG loci, especially during cell proliferation (197, 198). CpGs are distributed throughout the genome. CpGs can stand alone as an "isolate CpG" dinucleotides, or present clustered together in a CpG-rich regions called a CpG island, which are usually located in the promoter region of genes (199, 200). Methylation of CpG sites located in a gene promoter therefore often coincide with suppression of gene transcription as methylation of these sites tend to promote a closed conformation of the chromatin (201, 202). Indeed, Methylated DNA regulation of gene transcription process by either itself, physically blocking the binding of the TF (203, 204) or through the binding to proteins specifically recognizing methylated DNA such as Methyl-CpG-binding domain proteins and the DNA methyltransferases themselves (205, 206). The binding of these proteins recruits additional proteins such as histone deacetylases and other chromatin remodeling

proteins to the region, facilitating chromatin compaction and reinforcing a close chromatin structure forbidding the transcription process (205).

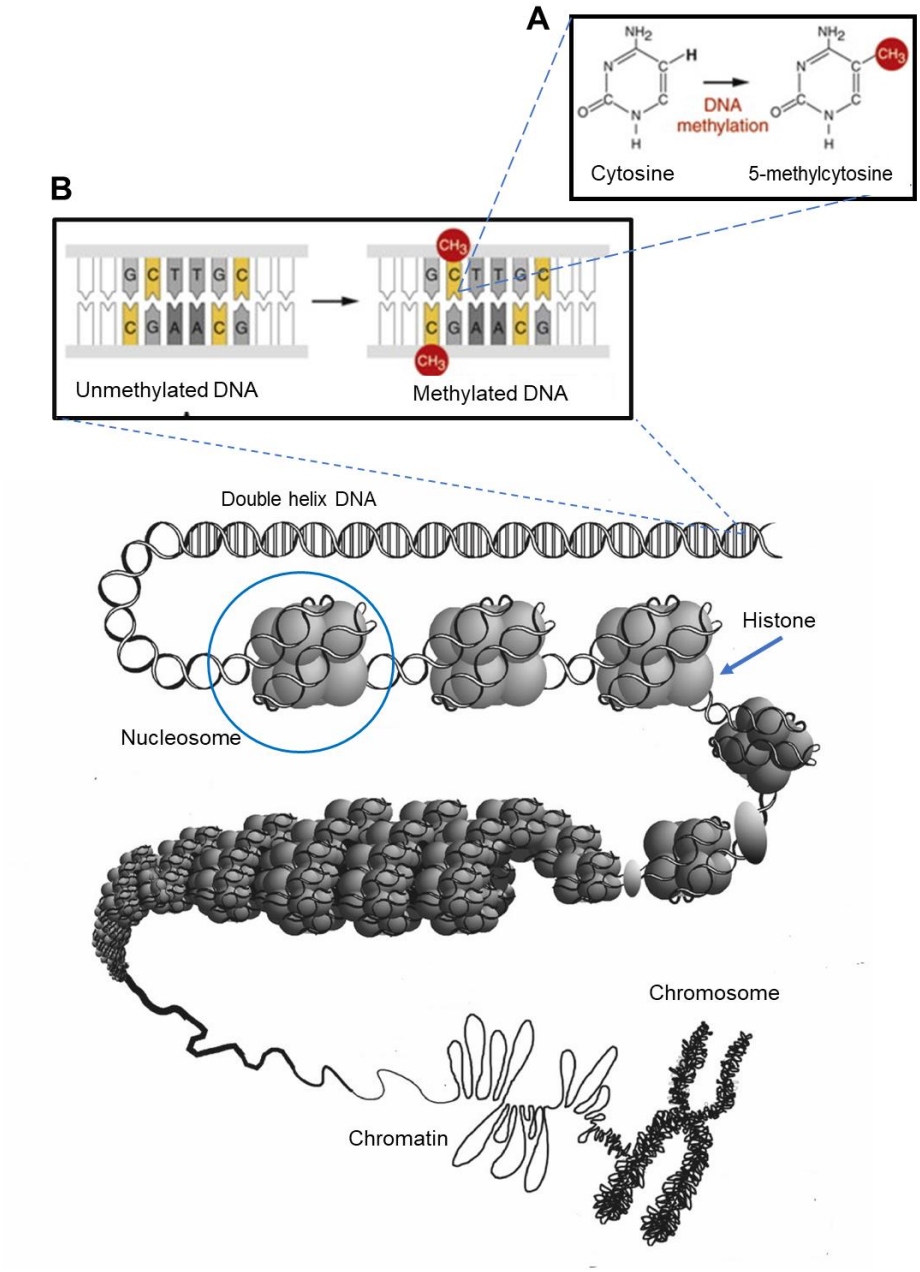


Figure 1-4 DNA packaging and DNA methylation.

A) The addition of a methyl (-CH3) group to DNA at 5' of cytosine base forming 5-methylcytosine (5-mC).
B) Methylation occur at cytosine-phosphate-guanine (CpG) position of double strand DNA.

(adapted from Molecular Biology, Chromatin Compaction, by Joëlle Brodeur and Martin Toussaint. 2006. DNA Methylation associated with diseases)

1.2.2 Methylation and diseases

Methylation plays a critical role in maintaining normal cellular processes including embryonic development, cell specificity, differentiation of cell, ageing, and many other processes (reviewed in (207)).

Abnormal patterns of methylation can lead to the pathological conditions including cancer, auto-immunity, metabolic disorders and many others. The effect of such abnormalities has largely been studied in cancer. On one hand, global **hypomethylation** of CpGs in tumour cells (often closely associated with repeated DNA elements) (208) disrupts the “normal” state of large regions of the chromatin, resulting in genomic instability and contributing to cell transformation. On the other hand, **hypermethylation** of CpG islands is also observed in a number of tumour suppressor genes promoter resulting in their silencing and the further contribution to transformation (188, 209, 210). Hypermethylation of *APC*, *RASSF1A* and *TP53*, promoter region has been reported as a common marker for early detection of cancer or evaluation of cancer development (211).

The original “injury” associated with such epigenetic modifications is still unknown for many diseases although environmental factors are likely to be causal particular in malignancies (173). Certain factors that have long been associated with cancer (for example carcinogens such as arsenite, chlorobenzene, nickel) do not have mutagenic abilities (as observed for other class of molecules that can directly alter the genetic code), however, they were shown to exert their effect through epigenetic mechanisms (212-214). These reagents may also be able to target the epigenetic machinery itself. Such epigenetic modifications are therefore now called epimutation and are as important in tumorigenesis as genetic mutations (173, 187, 210, 215). Punctual mutations are also not likely to cause major functional disruption if they are not directly associated with a mechanism allowing the development of the cancer (or other diseases). On the other hand, methylated cytosine create a “lesion” that cannot easily be discriminated by the mismatch mutation DNA repair mechanisms and may result in an increase G:T transition (i.e. a mutation)(216, 217), while, genetic mutations also often target genes implicated in the epigenetic machinery (218, 219). It was then suggested that an earlier involvement of epimutations and epigenetic mechanisms may allow DNA damages to persist unchecked by lowering the “proof reading” capability of cells (220). DNA damage accumulating over time would then lead to further increase mutations and vice versa. Another essential observation was then, the discovery of epimutations present in the tissue surrounding tumours (221, 222). This therefore suggested a wider local perturbation reflecting an initial epigenetic injury at a wider levels. A localised genetic event (potentially in a single cells) may then developed in this

predisposed tissue and turn into a transformed clone of cells and then into a cancer.

These observations led to the conclusion that mutations may not be the early events leading to cell transformation but a later consequence of an overall genetic instability initially created by epimutations. The nature of the original epigenetic injury is still elusive in certain forms of cancer, but in diseases like autoimmune disorders and particularly RA, inflammation appears a likely candidate while genetic events points to T-cells.

DNA Methylation in RA

In RA, there is increasing evidence showing alterations in methylation patterns. These modifications have been studied in SF, peripheral blood mononuclear cells (PBMC), and different subsets of immune cells and reviewed in (173, 185, 223).

Early studies showed that

- SF showed global hypomethylation, causing overexpression of inflammatory cytokines (224-226) as well as aberrant re-expression of LINE-1 elements (175, 176) a retrotransposon that contributes to the general genome instability. Overexpression of IL-6, an inflammatory cytokine that stimulates many inflammatory responses, was further specifically associated with the hypomethylation of CpG Island in its promoter (227, 228).

A few studies performed a close analysis of the change in methylation patterns in different immune subsets. In general, it appears that hypomethylation affected genes that are activated during immune responses, hence likely towards perpetuating the disease.

In T-cell,

- demethylation of the *CD40L* gene, crucial for immune activation of B-cells was observed, (229-231).
- hypermethylation in *CTLA-4*, protein receptor that functions as an immune checkpoint an essential molecule for Treg function (232, 233) suggesting impaired Treg, which play a role in suppressing the immune system
- decrease in methylation of the *FoxP3* gene, (the master regulator for the development and suppressive function of Treg) was associated with therapeutic effect of MTX (234), resulting in Foxp3 upregulation and consequently increase CTLA-4 normalization of RA Treg function.

In addition to the genes involved in immune response,

- various genes involved in the methylation and demethylation machinery itself (such as *DNMT1*, *TET1*, *TET2* and *TET3*) have also been reported to be altered at the DNA methylation levels in RA, which could have a major direct effect on the methylation status of more genes in T-cells (235).

These initial observations clearly suggested that methylation change could be involved in RA pathogenesis. The advance in microarray technology later facilitated the study of DNA methylation changes associated with RA genome-wide.

- 383 hyper- and 785 hypo-methylated CpGs were reported on CD4+T-cell from long lasting RA patients. These candidate CpGs were related to genes particularly involved in transcript alternative splicing and protein modification in addition to methylation changes in *HLA* genes (e.g. *HLA-DRB6, HLA-DQA1* and *HLA-E*) and other genes were also highlighted including *HDAC4, NXN, TBCD*,*TMEM61, ITIH3, TCN2, PRDM16, SLC1A5* and *GALNT9* (236).
- A second Genome-wide DNA methylation study again in long lasting RA patients from another group, identified 509 and 252 CpGs in T- and B-lymphocytes, respectively (237). *ARSB, DUSP22, GALNT9* were highlighted in T-cells and *ADAMTS17, ASB1, BARX2* in B-cells. They also emphasized the distinct changes in methylation pattern specifically associated with these two cells population in RA.

Most of the studies of DNA methylation in RA have focused on established RA, where different courses of disease progression, different treatments and the accumulated burden of chronic inflammation will have contributed to altered DNA methylation patterns while generating much heterogeneity at the same time. DNA methylation changes at the early stage of RA, while still naïve for disease modifying drug, have then gained attention.

- studies from SF, circulating T-cell, and B-cell in the early stages of RA demonstrated that DNA methylation alteration could occur early in the course of the disease and contributes to its development (238-240). The genes highlighted were related to the wnt signaling pathway for SF, DNA binding protein and transcription factor in T-lymphocytes, and kinase activity and plasma cell functions in B-lymphocytes study that compared DNA methylation change in early and established RA in T-cells showed
    - ~50 CpGs that were common between the two stages
    - 218 CpGs that are dominant in one of the stage.

Methylation changes that were dominant at the established stage are more likely to be the consequence of the progression as well as reflect the variations in the disease course and its response to various treatment over time, whereas methylation changes at the early (drug naïve) stage are more likely to be important to the initiation of the disease and contribute to its pathogenesis.

1.2.3 The role of Inflammation in triggering epigenetic changes

Epigenetic could be influenced by environmental exposures. Various factors such as air pollution, tobacco smoke, malnutrition, and all sources of stress could induce epigenetic change leading the pathological condition (241). Several pieces of evidence from oncology research showed that inflammation can contribute to tumorigenesis via alteration of epigenetic profile of cytokine genes for example. IL1β was reported to increases the risk of gastric cancer via the induction of aberrant DNA methylation profile in a mouse model (242). IL-6-induced inflammation was shown to promotes tumorigenesis in the oral cavity by altering LINE-1 element's methylation (243). The proposed mechanism was that IL6 regulates the expression of the *DNMT* genes, which encoding the methyltransferase enzymes responsible for establishing and maintaining DNA methylation, thereby regulating DNA methylation of other genes (244) and contributing to genetic instability and cancer progression. There is also other evidence (although with no clear mechanism) suggesting that oxidative stress and other pro-inflammatory cytokines (e.g. TNF-α and IFN-γ) could also induced methylation changes in cancer (review in (245)).

A study in RA also supports a role for inflammation-induced DNA methylation changes (246). It showed that chronic exposure to IL1 potentially contributes to global hypomethylation in RA SF also through the regulation of DNMT expression.

Altogether inflammation could be an important factor in driving aberrant DNA methylation, contributing to the progression of disease. Smoking, bacteria or virus infections and injury are well-known environmental risk factors for RA. It is possible that the inflammation caused by exposure to these factors triggers epigenetics change which together with RA genetic susceptible loci may be leading to pathogenesis.

1.3 The study of DNA methylation

There are several methods to study DNA methylation. The methods of choice in my project were adapted to the purpose of each part of my study towards

- the discovery of novel epigenetic change as an exploratory study of DNA methylation in different cell subsets from early drug naive RA patients
- the determination of regions/genes of interest with specific alteration in DNA methylation patterns.
- the analysis of a specific mark in more details towards a clinical applications (i.e. as a biomarker)

For an exploratory study, the recent advances in technology for next-generation sequencing and microarrays have allowed the epigenetic study of the whole genome with high resolution from a large number of samples. The technology that has been the most widely used for DNA methylation is genome-wide microarray as it provides a cost-effective platform and is easy to use experimentally (247-249). Bisulfite conversion of unmethylated cytosine to uracil on DNA chains helps distinguish between methylated and un-methylated cytosine while the Microarray technology enables the study of multiple specific CpG sites spread all over the genome.

The technology that was used to determine DNA methylation in samples in my project is the Illumina Infinium Human Methylation 450 Bead Chips array (Illumina Inc., CA, USA)(250, 251). It allows to assess methylation levels in more than 485,000 CpG sites per sample at a single-nucleotide CpG levels of resolution. It is designed to cover 99% of all known genes (by RefSeq) and 96% of identified CpG islands, with additional coverage for the regions flanking them (called shores/shelves) (251-253). This technology allows a wide view of the genome methylation status to make sense in biological terms and offers the opportunity to find all the methylation marks that are linked to a disease for example.

Computational tools are essential for handling the large and complicated datasets generated by microarrays. A limited range of publications describing computational analytical tools and packages were available in the public domain for the analysis of methylation microarray-derived data particularly beyond the initial stage of technical processing of the data and notably at the time I started my PhD. The programming language R is one of the most powerful open source tool for data analysis and statistical computing. The capabilities offered by R are extensive, particularly when used in combination with packages from Bioconductor (254, 255), the open-source software repository that provides tools

for the analysis and comprehension of high-throughput genomic data. Nonetheless, in order to get a biologically meaningful result, which highly depends on the research question, a specific strategy is needed to develop an analysis pipeline after the basic tools for pre-data processing are applied.

In contrast to genome wide array, the assessment of methylation at a specific site of interest, can employ several techniques. These are either based on the enzymatic reaction or DNA bisulfite conversion, then followed by downstream molecular techniques such as PCR and sequencing, pyrosequencing (256, 257), methyl specific PCR (258), quantitative Methylation-Specific PCR (qMSP) (259, 260), or Methylation-Sensitive High Resolution Melting (261). Which method to choose depends on the amount and quality of the DNA sample, the sensitivity and specificity requirements of the study, the robustness and simplicity of the method, the availability of specialized equipment and reagents, and bioinformatics software, and cost (248). The method of choice in my project, is bisulfite sequencing and qMSP which will be described in more detail in the relevant chapter.

Chapter 2 Overall Rational and hypothesis

The early pathogenesis of RA is still not clear and the series of events leading to disease remain to be elucidated as well as timed with respect to each other. In RA management, the diagnosis and treatment are not 100% effective. Understanding How and When genes/pathways involved in RA disease progression, will provide novel understanding about the pathogenesis which may be beneficial in many aspects including the discovery of new targets for treatment and for biomarker development.

DNA Methylation alteration are now central to new hypothesis about their role in leading to disease rather than being a consequence of genetic instability (alteration being detected before other molecular events). As DNA methylation can be influenced by the environment, modification detected could provide the missing link between such triggers and the resulting measurable events leading to disease. Methylation change detected in RA so far were shown to be associated with known pathological pathways, suggesting that they are indeed truly reflecting the disease situation and could provide more understanding of RA pathogenesis.

Considering that many RA susceptible loci are related to T-cell biological functions, all the other supporting evidences described for an pivotal role of T-cells in the initiating phase of RA progression, the importance of aberration in T-cell subsets at the disease initial stage (including the biomarker values associated with such disturbances), I hypothesised that methylation change in T-cells may occur early and can drive the disease towards chronicity.

Therefore a study of DNA methylation in T-cells at the early stages of RA (drug naïve) was initiated by my supervisor. CD4+T-cells comprise several subsets, naïve and memory, regulatory, polarised helper cells … etc, which are different in nature and importantly use epigenetic changes as part of their differentiation commitment. Considering that methylation change associated with the disease should be distinct from the epigenetic changes resulting from the differentiation of these cell subtypes, there could be more potential for discovery in taking a closer look at naive CD4+T-cells.

The work performed over 20 years by my supervisor also pointed to changes and/or defect in naïve CD4+ T-cells (48, 92, 94, 95). This study, therefore used an illumina methylation genome-wide technology, for an exploratory study of RA pathogenesis, measuring DNA methylation of early, drug naive RA compared to HC, in naïve and memory CD4+ T-cells. Monocytes were also included as a

reference for another immune cell that respond to inflammation stimuli and also has an important role in RA.

The aim of my project is first to identify changes in DNA methylation in these cells in early RA, to understand more about RA pathogenesis and second to select a CpG target for biomarker development.

My thesis is therefore based on a general hypothesis and one main assumptions:

General hypothesis

- Change in DNA methylation in CD4+ T-cells happen early in RA and may drive the disease development by altering important physiological pathways.

Assumption

- Naïve CD4+T-cells are the main cells targeted by such changes

My project is organised in 2 main part

1. gain more understanding of early events/pathways in disease pathology by studying genome wide data on DNA methylation
2. select potential CpG candidates for the development of a biomarker for the prediction of clinical outcome using the qMSP technic.

Chapter 3 Materials and Method

This chapter describes the general material and method used in my thesis. All detail of reagents, kits, buffer and instruments used in this research can be found in Appendix 1.

3.1 Ethics, Patients

My project used samples from healthy volunteers and patients who were enrolled from an early arthritis clinic (EAC) into an observational register (named IACON/RADAR) of the Inflammatory Arthritis disease continuum at the Chapel Allerton (ChA) hospital in Leeds. All subjects and healthy volunteers gave informed written consent. The observational IACON/RADAR study obtained ethical permission in 2010 under the REC number: 09/H1307/98 (ethical approval letter attached in Appendix 2).

In EAC, patients were examined using various clinical assessments (joint counts notably), diagnostic blood tests (for autoantibodies, inflammation markers), medical history, physical questionnaires of well-being, patient ability to perform daily tasks and imaging (ultra sound) in order to be classified toward a specific type of arthritis and to receive proper treatment. These parameters at the first visit (referred to as baseline) and follow-up visits (3, 6 and 9 months and yearly) were recorded in databases. Blood samples were collected, processed, and stored by the ChA tissue bank team for research purposes at baseline and 6 months.

At the first visit, some patients expressed features that met the classification criteria for RA or another type of arthritis (psoriatic, reactive, or osteoarthritis), another type of inflammation (gout, connective tissue inflammation) and can be allocated a specific diagnosis, while others remain unclassified (i.e undifferentiated arthritis, UA). Some patients also spontaneously resolve symptoms and are discharged. In some cases, it took up to 2-3 years to be able to be diagnose RA.

For this project, I used different

- sample groups : HC, RA, other arthritis or UA,
- sample types : fresh blood, serum, frozen PBMC, frozen whole blood
- time points : baseline or week 24

in different experimental designs. The demographic and clinical details of the samples used in the different lines of work are described in the corresponding results chapters.

3.2 Analytical resources

My project needs access to several types of analytical tools and resources. Most of the bioinformatics work was performed with the R programming languages (262) and R packages available in Bioconductor (255), an open source software providing tools for the analysis and comprehension of high-throughput genomic data. The R packages and other analytical tools used in my project are listed in Table 3-1.

Table 3-1 Analytical tools and data resource for the analysis

Analytical tools		
R Package	**Package source and description**	**Package web link**
dplyr (263)	R package for data manipulation	https://CRAN.R-project.org/package=dplyr
reshape2 (264)	R package for data restructure and aggregation	https://CRAN.R-project.org/package=reshape2
rms (265)	R package for Regression model strategies	https://CRAN.R-project.org/package=rms
qqman (266)	R package for generate Manhattan plot	https://cran.r-project.org/web/packages/qqman/ index.html
gplots (267)	R package for data visualisation; Heat maps/ hierarchical clustering	https://CRAN.R-project.org/package=gplots
ggplot2 (268)	R package for data visualisation	https://cran.r-project.org/web/packages/ggplot2/index.html
Minfi (269)	R/Bioconductor package for DNA methylation data pre-processing, analysis and visualisation	http://bioconductor.org/packages/release/bioc/html/minfi.html
FDb.Infinium Methylation. hg19 (270)	R/Bioconductor package for annotation Illumina Infinium DNA methylation probes	http://bioconductor.org/packages/FDb.Infinium Methylation.hg19/
genefilter (271)	R/Bioconductor package for statistical analysis from high-throughput experiments	https://www.bioconductor.org/packages/release/bioc/html/genefilter.html
bumphunter (272)	R package for data Differentially methylated region (DMR) finding	https://www.bioconductor.org/packages/release/bioc/html/bumphunter.html
DMRcate (273)	R package for data DMR finding	http://bioconductor.org/packages/release/bioc/html/DMRcate.html
Web-based tool		
Bio venn (274)	web-based tool for comparison and visualization of biological lists using area-proportional Venn diagrams	http://www.biovenn.nl/
Panther (275)	web-based tool for gene classification	http://pantherdb.org/

String network analysis (276)	web-based tool and database of known and predicted protein-protein interactions	https://string-db.org/
Other Software		
Primer Express™ Software v3.0.1 (Applied Biosystems™)	Primer design software	Applied Biosystems™
Methyl Primer Express™ Software v1.0	Primer design software	Applied Biosystems™
MethPrimer 2.0	Primer design software	http://www.urogene.org/methprimer2
Bi search web server	Primer design software	http://bisearch.enzim.hu/
Sequencing Analysis 5.2 (Applied Biosystems),	Sequencing analysis software	Applied Biosystems™
The Design and Analysis II	qPCR analysis software	Themo fisher could application.
Data resource		
UCSC database	Genome browser	UCSC Genome Browser: Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, Haussler D. The human genome browser at UCSC. Genome Res. 2002 Jun;12(6):996-1006.
NCBI GEO dataset	database for methylation data and gene expression data	https://www.ncbi.nlm.nih.gov/gds

3.3 Initial procedures for obtaining Genome-wide DNA methylation data: Samples procedure and data acquisition

Before I started my PhD, a Genome-wide DNA methylation array (Illumina Infinium Human Methylation 450 Bead Chip) was performed by an external contractor (Hologics, Manchester) on peripheral naive CD4+T-cell, memory CD4+T-cells, and monocytes from 6 HC and 10 early patients (who meet the EULAR-2010 classification criteria for RA (1)), being naïve for disease-modifying anti-rheumatic drug (DMARD) and having an active disease with at least 3 swollen joints and raised inflammation markers (CRP>10 mg/L). The raw data were provided to my supervisor as 48 idat files. I performed an analysis of this data set.

The detail of samples processing, and data acquisition are described in the appendix 3. These were performed by a research technician before my arrival in Leeds.

3.4 Genome-wide DNA methylation data analysis

A data analysis pipeline was designed using a combination of R (V3.3.1)(262), Bioconductor R packages, in-house R scripts, and other analytical tools. An overall analysis and experimental workflow show in Figure 4-1in part 1 introduction.

3.4.1 Quality control and Data pre-processing

A total of 48 genome-wide DNA methylation profiles (from 3 cells subset of 6 HC and 10 early RA patients) were retrieved in IDAT files and were loaded into R. Each sample provided reads for signal intensities (for both methylated and unmethylated DNA) for the 485,512 CpG site (each specifically associated with a probe). Data quality control analysis and pre-processing were performed with the R package Minfi (269). Plots of Methylation levels (in form of β-values) for density including all 48 samples and bean plots for each individual sample were generated using the same R package.

CpG probes which were identified to be common SNPs and cross-reactive probes that have been shown to hybridise to multiple locations in the genomes were filtered out to prevent false interpretation of the methylation signal difference(277). Methylation levels for each CpG site were presented as β-value or M-value according to the analysis to be performed.

- β-values are the ratio of the fluorescence intensity between the methylated and unmethylated probes, ranging from 0 (all copies of the CpG in the sample are un-methylated) to 1 (all copies of the CpG in the sample are methylated).
- M-values are log-transformed β-values preferably used for statistical testing(278).

3.4.2 Initial exploration of data

3.4.2.1 Visualizing Multivariate Data

Multidimensional scaling (MDS) for (i) each cell types, (ii) genders, and (iii) RA versus HC was performed to examine the source of variation in the dataset and were plotted using the mdsplot function in the minfi package in R(269).

3.4.2.2 Identification of Differentially Methylated individual CpG site (T-test, Manhattan plot, heatmap)

2-sided t-tests (on M-value) were performed on every CpG using the function rowttest in the genefilter package (271) for significance of the difference in methylation between HC and RA.

A manhattan plot for illustrating the levels of significance of differentially methylated CpGs between HC and RA was plotted using -log10(P-values) for each probe on the array against its chromosomal position using R package “qqman’’ (266). Three thresholds defining high/medium/low level of significance for differentially methylated CpG were set at p-value ≤ 0.0001, ≤ 0.001, and ≤ 0.01, respectively.

A hierarchical clustering and a heatmap of the significant DM CpG (P-value ≤ 0.01) between HC and RA was generated for visualising the DNA methylation level and observing the relationship between groups (HC/RA). This was generated from β-values using the heatmap2 function of the gplot package (267).

3.4.2.3 Annotation of CpG Island and Gene information associated with individual CpG/probe

The relevant information related to CpG Island and gene associated with each probe (gene symbol), were retrieved before further analysis to select candidate CpGs with biological meaning. Each CpG sites/probe was related to annotation for CpG Island information on position, size, enrichment of CpG, number of probes in the island, and the associated gene using using getAnnotation in the minfi package (269) and the getNearestTSS function in FDb.InfiniumMethylation.hg19 package (270). It is worth noticing that this package associates the nearest Gene transcript to a CpG, however, this may not be its true biological association.

3.4.3 Developing tool to identify Differential Methylation (DM)

To identify clusters of DM CpG, a custom R scripts to score each individual CpG and prioritise them was developed (the concept and details will be described the results section, full code available on request).

3.4.4 Further analysis to understand the biological relevance of DM gene to RA pathogenesis.

Several tools were used to understand the relevance of DM. This included Gene Expression Omnibus (GEO) repository to retrieve publicly available gene expression data, and STRING database (279, 280) to examine the interaction the proteins associated with DM. I worked in collaboration with other students in my supervisor research group for some of this analysis. This will be clearly mentioned in each particular section of the result part.

3.5 Sample preparation

3.5.1 Isolation of peripheral blood mononuclear cells (PBMC) from human peripheral blood (Ficoll)

A sample of peripheral blood (PB) was collected from all participants in EDTA containing vacuette blood bottles. The PBMCs were isolated by density gradient separation using LymphoprepTM, according to the manufacturer's instructions. Briefly, PB was dilute 1:1 with PBS before slowly layer on to the lymphoprep (with ration 1 lymphoprep: 2 diluted PB) at RT and centrifuged at 2,400 rmp, with no break for 20 min at RT. The cloudy lymphocyte layer was then aspirated and transferred to new falcon tube containing cold PBS before centrifuge at 1,800 rpm, 10 min at 4^{o}C. The supernatant was discarded, and cell pellets were washed twice with cold PBS before centrifuging at 1500 rpm, for 10 min, at 4^{o}C. After removing the supernatant, cell pellets were kept on ice until ready to for further steps. Cell counting was performed during the washing step or after finish as described below.

Fresh PBMS samples were used for fluorescence-activated cell sorting.

3.5.2 Cell counting and viability Testing with Trypan Blue Exclusion Method

Cell number and viability was assessed by Trypan blue on the basis that living cells possess intact cell membranes that exclude the dye while dead cells do not. The cell suspension was mixed with 0.4% trypan blue dye in 1:1 dilution. Live/dead cells were counted with a haemocytometer with a light microscope.

3.5.3 Fluorescence-Activated Cell Sorting (FACS)

FACS is a technique that identifies and purifies a cell subset form a mixture of cell populations. It used the capability of flow cytometry to detect the specific light scattering characteristics of each cell and the fluorescent labelled attached to cell surface molecules by specific antibodies.

In this project, I sorted 5 cell subsets: CD4 T-cells, CD8 T-cells, NK cells, B-cells, and monocytes. In brief the, PBMC were isolated from fresh EDTA blood and stained with antibodies using a standard cell surface staining protocol. (see below).

3.5.4 Standard cell surface staining protocol

Cell pellets were resuspended in blocking buffer for 30 min at 4^{o}C. Blocking buffer was removed by centrifuging at 500g, for 5 mins at 4^{o}C (same for other centrifuge steps) and the supernatant was poured out. Cells were resuspended in 1x 10^{6} cells/100 µL of FAC buffer were stained with antibodies according to the purpose of the experiment (see detail below), for 30 min at 4^{o}C in the dark. Stained cells were then centrifuged to discard the excess antibodies and washed once with FACS buffer before resuspending in 200 µL of FACS buffer for analysis on the FACS machine.

For cell sorting, cell stained with a single cell surface marker (i.e., 1 antibody only) were used first for setting gates. Then cells stained with a mixture of all antibodies were sorted on the Influx FACS machine, operated by the Flow Cytometry and Imaging facility support staff. Antibodies used for staining each cell type and an example for gating are showed in Figure 3-1,A (antibodies detail are described in Appendix 4).

For purity (see below), cells were stained with a single antibody (anti-CD4) or together with anti-CD3 (optional) and analysed on Attune machine, against light scattering. Antibodies used for staining and an example for gating are showed in Figure 3-1, B.

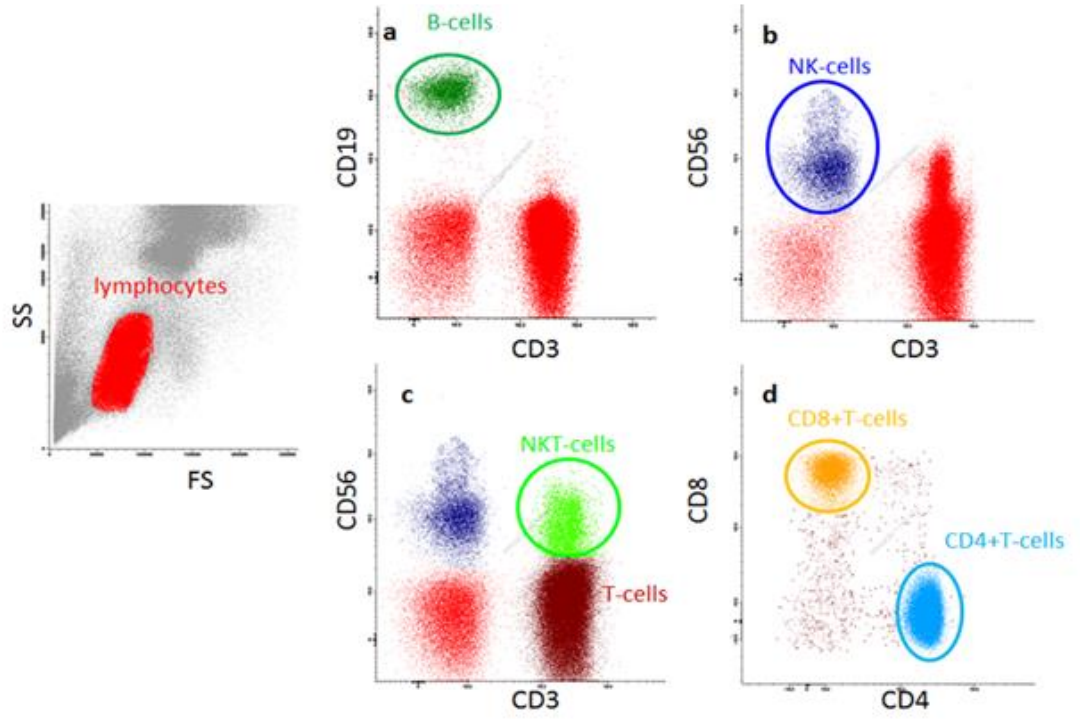
A: Cell sorting

Antibodies for cell sorting assay

Staining tube	Antibody	Dye	Volume used/ $1x10^6$ cells
Single staining 1	CD3+	FIT C	6 µL
Single staining 2	CD4+	APC-cy7	4 µL
Single staining 3	CD8+	A700	4 µL
Single staining 4	CD14+	Pacific Blue	6 µL
Single staining 5	CD19+	APC	10 µL
Single staining 6	CD56+	PE	10 µL
Mix staining	Combination of all 6 antibodies above		

Gating detail for cell sorting

Target cell	Antibody	Dye
CD4+T-cell	CD3+	FITC
	CD4+	APC-cy7
CD8+T-cell	CD3+	FIT C
	CD8+	A700
B-cell	CD19	APC
NK cell	CD56+ CD3 -	PE (no FITC)
NKT cells	CD56+ CD3+	PE FITC
Monocyte	CD14+	Pacific Blue



B: CD4+T-cells purity check

Antibodies for CD4+T-cells purity check

Staining tube	Antibody	Dye	Volume used/ $1x10^6$ cells
Single staining 1	CD4+	V500	5 μL
Single staining 2 (optional)	CD3+	FIT C	6 μL

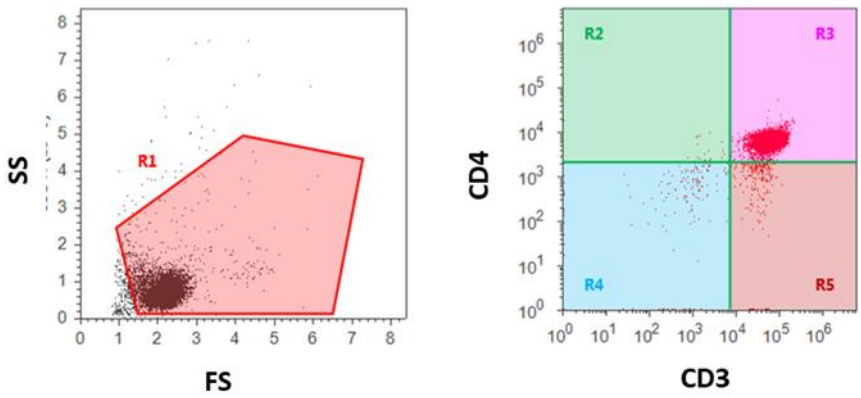


Figure 3-1 Cell sorting and CD4+T-cells purity check

A) Antibodies staining condition detail and gating for cell sorting and B) CD4+T-cells purity check assays.

3.5.5 Working from frozen PBMC

The Frozen PCMC were taken out from the -150^{o}C freezer ChA tissue bank. Cells were rapidly thawed by gently swirling the vial in the 37^{o}C water bath until only a small crystal was left in the vial. Cells were then transferred to a 15 ml tube and diluted with ice-cold PBS. Cells were maintained in a cold condition for all of the following step. The Cell suspension was centrifuged at 500g for 5 min at 4^{o}C before the freezing milieu was discarded. Cells were then washed with ice-cold PBS and centrifuged. Cell counting was performed after the final wash.

Of note, after thawing, frozen cells may be sticky and clumped as there was a lot of cell death and cell debris. Such PBMC could be used directly for DNA extraction. However, for CD4 T-cell isolation, well-separated cells are needed and the clumped PBMC cells needed to be discarded. Filtering aggregated suspensions through a 37 µm cell strainer or using a DNase treatment was tested as an alternative option.

3.5.6 CD4-T-cell isolation by magnetic bead and purification check by flow cytometer and purity check

CD4+ T-cells were isolated from frozen PBMC using the immuno-magnetic negative selection, Human CD4+ T Cell Isolation Kit (EasySep™). The principle is to label unwanted cells with antibody that will form complexes with magnetic particles before using a magnet to separate the magnetically labelled cell from the untouched target cells, here the CD4+T-cell. CD4+T-cell isolation was performed according to the kit protocol with optimisation for frozen cells. In short, PBMC were suspended at 5 x 10^{7} cells/mL in recommend buffer (PBS containing 2% FBS and 1 mM EDTA) adjusting volume depending on cell counts in polystyrene round-bottom tube. 50 µL/mL of the Isolation cocktail of antibodies were added to the cells and incubated for 5 min, at RT.

The magnetic beads (RapidSpheres™), were then added at 50 µL/mL to the sample. The sample was mixed and toped-up with the buffer to 2.5 ml before placing the tube into the magnet for 3 min to allow magnetically labelled cell to attach to the tube's wall. The non-labelled cells (CD4+T-cells) were then poured out into a new tube. After isolation, the purity of the isolated CD4+T-cells was checked by flow cytometry using a single antibody (anti-CD4) labelled with V500 (5 µL per test). Cells were stained according to the standard protocol described above and analysed as per the gate described in Figure 3-1. Cells were also counted as described above.

CD4+T-cells with purity more than 90% were used in further experiments. Cells were immediately processed for DNA isolation or stored as cell pellets at -80^{o}C for future use.

Notes;

- Contaminants were mainly unidentifiable debris (i.e. not other cell types) hence the needs to eliminate such samples as they would be containing large amount of DNA from unknown cells.
- In case of having a limited number of target cells, after pouring the CD4+T-cells fraction out, the tube with labelled-unwanted cells on the wall was removed from the magnet and were resuspended with 2.5 mL recommend buffer. The tube was then placed again into the magnet to gain more CD4 T cells that might stayed in the bottom of the tube. This did not affect the purity of isolated cells, but gained more CD4+T-cells.

3.5.7 DNA isolation and quality check

Genomic DNA of target cells (from both fresh or frozen cells) were extracted using a silica-membrane-based DNA purification principle (QIAamp DNA Blood Mini Kit). DNA isolation was performed according to the manufacturer's protocol. Briefly, Cell pellets, up to 5 x 10^6 cells resuspended in in 200 µL PBS were added to 20 µl of Proteinase K, followed by 200 µL of lysis buffer (Buffer AL) in a microcentrifuge tube and incubated at 56^{o}C for 10 min. DNA was precipitated by adding and mixing well with 200 µl of 100% ethanol. The sample was transferred to a QIAamp Mini spin column, then centrifuged to separate the supernatant at 15000 g for 1 min (used for all centrifugations). The DNA which is adsorbed onto the QIAamp silica membrane in the column was washed using 2 centrifugation steps with 2 different wash buffers, (Buffer AW1 and Buffer AW2), to remove any residual contaminants. Purified DNA was then incubated with Buffer AE for 10 min before being centrifugated to elute the DNAfrom the QIAamp Mini spin column.

DNA quality and concentration were asserted using a spectrophotometer (ND1000). An absorbance ratio (OD280/OD260) at 1.8 – 2 was accepted as pure DNA and such samples were used for further experiment. Poor quality DNA was also used although data were analysed separately to ensure alignment with good quality samples. Isolated DNA was stored at 4^{o}C for use within 1-2 week or at -20^{o}C for long-term storage.

3.6 ELISA

Frozen serum samples were retrieved our departmental tissue bank from -80°C freezers. Two commercial Sandwich enzyme-linked immunosorbent assays (ELISA) kits were used to measure cytokines: IL-21 and IL-34 (Biolegend, UK). ELISA was performed according to the manufacturer's instructions. Briefly, 50 µL of the standard or sample was added pre-coated plated with a monoclonal mouse anti-the specific human antibody and incubated while shaking at 200 rpm for 2h. The plate was washed 4 times with wash buffer before adding 100 µLof the Human detection antibody solution and incubate while shaking for 1h. The sample was washed and then incubated with 100 µL of Avidin-HRP for 30 min followed the wash and 100 µL of substrate solution for 10 min. 100 µL of stop solution was added in the final step and the absorbance was read at 450nM. The cytokine concentration was calculated by comparing the absorbance to the standard curve in the log-log graph. Non-parametric Mann-Whitney U-test was performed on data comparing HC and RA. Statistical analysis was performed in SPSS V24.

3.7 Bisulfite sequencing

3.7.1 Method Principle of bisulfite sequencing

Bisulfite sequencing is considered the gold standard technique used to study DNA methylation in a region of interest at a single nucleotide resolution (281). It is the combination of two techniques: bisulfite conversion and sequencing. Bisulfite conversion help distinguish the unmethylated from the methylated cytosines. The sequencing technology provided effective access to the nucleic acid sequence.

Bisulfite sequencing includes 3 main steps:

- bisulfite conversion,
- PCR amplification,
- sequencing and methylation status analysis.

Bisulfite conversion

The conversion of cytosine using Bisulfite makes it possible to map un-methylated cytosine as these are change into a different base while methylated Cytosine (in a CpG site) are protected from that conversion (282). The reaction of bisulfite conversion involved denaturation of genomic DNA, deamination of unmethylated cytosine residues and desulfonation (Figure 3-2). This treatment causes the unmethylated cytosines to convert into uracil, which will then be amplified as a thymine in a following PCR, while methylated cytosines remain intact (Figure 3-3). Conversion allows to distinguish between methylated and unmethylated-cytosine in a base level, enabling DNA methylation studies.

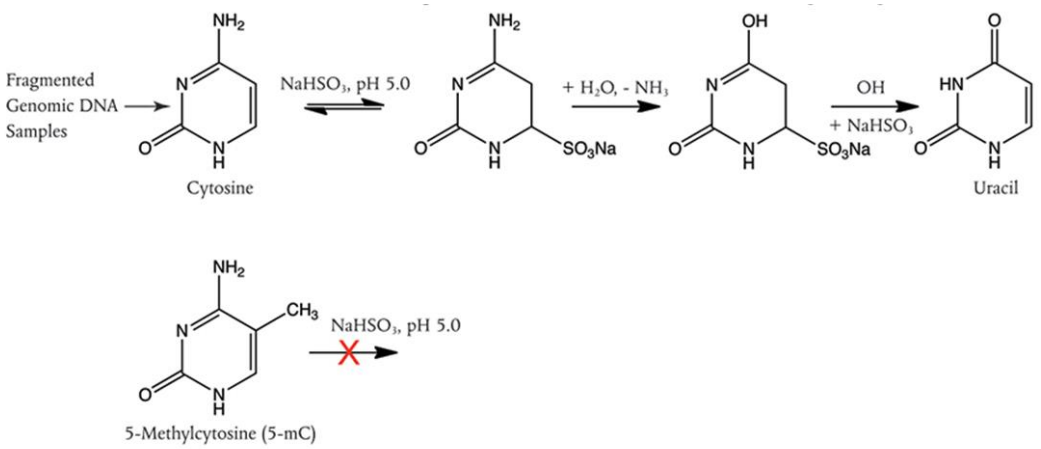


Figure 3-2 The chemical reaction that underlies the bisulfite conversion of cytosine into uracil. (modified from (283)) Genomic DNA is denatured and treated with sodium bisulfite (sodium hydrogen sulfite NaHSO3,) leading to the conversion of cytosines tocytosine sulfonate, which is converted touracil sulfonate, and then desulphonated to uracil while the methylated cytosines remain intact.

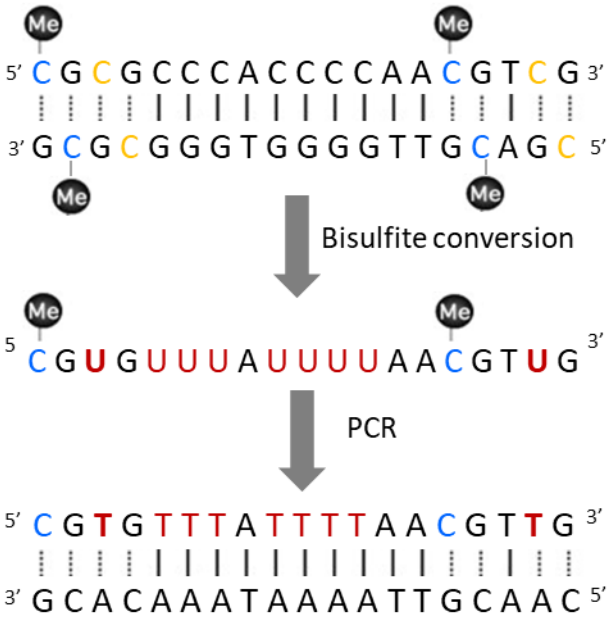


Figure 3-3 DNA sequence after Bisulfite conversion and PCR amplification. The free cytosine (shown in black) and unmethylated cytosine at CpG site (shown in orange) are converted to uracils by bisulfite and changed to thymine after PCR reaction. In contrast, methylated cytosines at CpG sites (shown in blue) remain cytosines after the conversion and PCR.

PCR amplification

A second important step is PCR amplification of the target region from the bisulfite-converted DNA. In this step, the target region should be amplified regardless of methylation status. Therefore, primers should be designed in a region that do not contain CpG site. It is however difficult to achieve this as DNA methylation studies when looking at CpG rich region or CpG island. If unable to avoid any CpG site, the CpG site should then be located at the 5' end of the primer where the mixed bases (either cytosine and thymine) should not interfere too much with priming and allowing DNA amplification.

Primer design guideline for bisulfited converted DNA

Unlike normal PCR, designing the PCR reaction for bisulfite converted template is challenging. Conversion of many unmethylated cytosines to thymine reduces DNA complexity. Furthermore, as a result of the conversion, DNA remains relatively denatured by high temperature (i.e. no longer fully complementary double strand). DNA also becomes highly fragmented.

Primer design is critical for a successful amplification and thus should be done considered these facts.

- As DNA is not complementary, primers for the sense strand are different from the anti-sense strand. The methylation status of CpG could be accessed using primer design for either of them. The lower complexity of the DNA increase the possibility of promiscuous binding, making it difficult to design PCR with high specificity. Primer, therefore, need to be longer (26-30 bases) than normal PCR and the melting temperature should be higher (>55^{o}C) than for PCR (and even higher for qPCR, see later). In the complete bisulfite treatment reaction, all free cytosines (cytosine at non-CpG site) are converted to thymine. Thus, the free cytosine should be included in primer sequences to ensure amplifying of the fully converted DNA.
- Fragmented template DNA (by the bisulfite treatment) limit the length of PCR product. Primers should aim for the amplicon between 150-300 bp.

In addition, designing bisulfite converted PCR primers should also follow the basic rule of PCR, For example,

- the forward and reverse primers should have a compatible melting temperature,
- they should not form duplexes with each other or themself,

- they should be specific for the target region, and not for other regions in the genome.

Aside from primer design, PCR condition could also contribute to the successful amplification. It is recommended to use

- Hot Start polymerase to improve the reaction specificity.
- Because DNA is no longer complementary Increase the PCR cycle to 35 - 40 cycles (it takes longer before the reverse primer can bind to its complementary template which is generated just from the forward primer.

Sequencing and methylation status analysis.

The third step is the sequencing of the amplified target region. The sequencing remains based on the dideoxynucleotide chain termination and capillary electrophoresis-based separation of the sequencing products. The results are displayed as an electropherogram in which a sequence of peaks each representing a base-position and the four colours identifying the base (A, T, C or G).

The DNA methylation status at each CpG site can be interpreted by analysing the electropherogram peak. The CpG site (after conversion) can be observed by reference to its position in the original DNA sequence (unconverted). Unmethylated free cytosines are converted to thymine while methylated cytosines in CpG remain the same or not if the CpG is un-methylated. At the individual CpG site, the presence of a C-peak therefore indicates a methylation at CpG site, the presence of a T-peak indicates an un-methylation CpG site, the presence of a mixture of both C and T-peaks indicates partial methylation at that CpG position.

It is important to consider that at each individual CpG site in the genome, a cytosine can either be methylated or unmethylated. DNA samples are derived from the DNA of populations of cells, some that could have methylated CpG, and some cells that could have unmethylated at each CpG site. Therefore, the methylation status of individual CpG refers to the proportion (hence usually described as a percentage) of cell in that population in which cytosine are methylation at that CpG site.

3.7.2 Method detail of bisulfite sequencing

3.7.2.1 Bisulfite conversion

To differentiate between unmethylated and methylated cytosine in DNA methylation studies, Bisulfite conversion was performed using the EZ DNA Methylation-Gold Kit following the manufacturer's protocol. In short, 500 ng of genomic DNA in 20 µL PBS and 130 µL of the CT conversion reagent were mixed to a final volume of 150 µL in PCR tubes and placed into a Gradient Thermal Cycler programmed to denature at 98 ºC and incubate at 64ºC for 10 min and 2.5 h, respectively. Samples were then transferred to the Zymo-Spin™ IC Column for desulphonation and clean-up through adding 200 µL of desulphonation buffer followed by several washing and centrifugation steps (15000 g for 30 at RT). Pure Bisulfite-converted DNA was eluted from the column matrix with 20 µL of Elution Buffer and used for immediate analysis for PCR/sequencing or stored at -20ºC for later use. Note: Bisulfite converted DNA is less stable, it is better to use within 1-2 week. The Quantity of bisulfite converted DNA was measured by nanodrop using the value of 50 µg/mL for Ab260nm =1.0)

3.7.2.2 Direct Bisulfite sequencing

Bisulfite sequencing was used to access DNA methylation status at target CpG sites of the candidate genes.

3.7.2.2.1 Primer design for Bisulfite-converted DNA

Primers were designed to amplify a region containing the candidate CpG sites to be tested for methylation status from Bisulfite converted Genomic DNA according to the primer design guideline. Design could use either sense or antisense DNA strands.

First, the genomic DNA sequence of candidate regions was obtained from UCSC or NCBI database (Human Genome Assembly GRCh37.p13). The sequence was then bisulfite converted *in silico,* using the sequence manipulating function in the online platform, MethPrimer 2.0. Forward and Reverse primers were designed using 3 primer design softwares (MethPrimer 2.0, Bisearch, and Methyl Primer Express™ Software v1.0). The primer sequences were modified to obtain the suitable Tm (between 55 and 60 ºC) and to avoid self-dimers/hairpins or primer-dimers.

The PCR product was checked *in silico* to ensure the specificity to the genome, absence of homology with other sequences of bisulfite-converted human genome DNA, using the blasting function of the Bisearch primer-design and search tool.

All primers were synthesised by Thermo Fisher Scientific, reconstituted to a concentration of 100 uM in sterile water. Primers were aliquoted stored at -20^{o}C. The detail of each primer described in the appropriate result sections.

3.7.2.2.2 PCR Amplification of target DNA

The target DNA regions containing the candidate CpGs was amplified from Bisulfite converted DNA by PCR with the specific-design primer pair. In order to obtain the amplified product, PCR conditions were optimised from a standard PCR mixture and conditions as recommend by the manufacturer's instructions of the HotStarTaq DNA Polymerase and HotStarTaq Master Mix Kit and described in Table 3-2. All reactions were performed in the TC-512 gradient thermocycler machine. The PCR product size and primer specificity were analysed by Agarose gel electrophoresis and visualized under the UV light (see agarose gel section).

In order to ensure the specificity and a good yield of each PCR product for each primer pair, the following factor were optimised.

- Aneling temperature,
- Mg2+ concentration,
- Primer concentration (labelled in blue in Table 3-2).

Successfully amplified PCR products were used immediately in the next step or were stored at 4^{o}C for use within 2-3 days.

Table 3-2 Composition of PCR reaction and PCR cycling program recommended by the manufacturer

A : Composition of PCR reaction recommended by the manufacturer

Stock conc.	Reagent	Final conc.	vol/1 reaction
10X	PCR Buffer*	1x	2
25mM	$MgCl_2$	**1.5 mM (vary at 1.5-3 mM)**	0*
10 mM of each	dNTP mix	200 uM of each	0.4
100 uM	Primer A	**0.5 uM (vary 0.1-0.5 uM)**	0.1
100 uM	Primer B	**0.5 uM (vary 0.1-0.5 uM)**	0.1
5 Unit/ul	HotStartTaq DNA polymerase	2.5 U/reaction	0.5
	Distilled water		14.9
	Template DNA	< 1ug/100 ul reaction	2
	Total Volumn		20

*Buffer contain 1.5 mM

B: PCR cycling program recommended by the manufacturer

PCR cycle			
Initial activation	95 °C	15 min	
Denaturation	94 °C	0.5-1m	25-30 cycles
Annealing	50-68 °C	0.5-1m	
Extension	72 °C	1 m	
Final extension	72 °C	10 min	

3.7.2.2.3 Agarose gel electrophoresis

The PCR products were analysed by Agarose gel electrophoresis. 1.5% w/v of Agarose gel was prepared in Tris-borate-EDTA buffer (TBE) and 5µl ethidium bromide /100 ml of Agarose gel. 5 µL of PCR product and 1 µL of 6X gel loading dye were mixed and loaded into a well of the Agarose gel. 50 bp or 100 bp DNA ladder were used as size marker. The agarose gel electrophoresis was run at 100 V for 20-45 min in 1x TBE buffer and visualized the under the UV light (ChemiDoc Imaging Systems,).

3.7.2.2.4 Sequencing

Purification of the PCR product

The amplified PCR product (amplicon) from converted DNA was purified prior to performing the direct sequencing in order to remove unincorporated primers and dNTPs that might interfere with the sequencing result. 2.5 µL of PCR product was added with 1 µL of ExoProStar enzyme, which include Exonuclease I and Alkaline Phosphatase, and incubated at 37^{o}C, for 15 min, and then at 80^{o}C, for 15 min for enzyme inactivation.

Sequencing reaction

The purified PCR amplicon was used as the template for sequencing using BigDye™ Terminator v3.1 Cycle Sequencing Kit. The mixture and thermocycling condition of the standard reaction was described in Table 3-3. The sequencing reaction for each target region was performed according to the standard recommendation of the manufacturer's protocol. The reaction was performed in TC-512 thermocycler machine.

Before loading the sequencing reaction onto the analyzer, the product of the sequencing reaction was cleaned-up using ethanol precipitation. Briefly, 10 µL of the reaction product was added with 3 µL sodium acetate (3M, pH 5.2) and 30 µL 95% ethanol. After incubation at RT for 30 min, the DNA was then centrifuged at 2254 g, for 30 min. DNA pellets were washed with 70% ethanol and dried for 1 minute at 95°C. The purified pellets were re-suspended in 20 µL of HiDi formamide and sequenced on the 3130xl Genetic Analyzer.

Table 3-3 Composition of sequencing reaction and thermo cycling program recommended by the manufacturer

A : Composition of sequencing reaction

Reagent	vol/1 reaction (uL)
Ready Reaction Mix (RRM)	4 ul
Primer 3.2 uM (forward only)	1 ul
Nuclease free water	4 ul
DNA Template	1 ul
Total volume	10 ul

B: Thermo cycling program

Thermo Cycle			
Initial denaturation	96ºC	1min	
Denaturation	96ºC	10 sec	28 cycles
Annealing	50ºC	5 sec	
Extension	60ºC	4 min	
	Hold at 15ºC		

To obtain the best quality sequences for specific PCR product, optimisation was performed on these following parameters:

- DNA template concentration,
- Primer concentration,
- RRM enzyme concentration,
- thermocycling conditions.

The controls used to optimize each step prior to use the patient DNA samples were :

- fully Methylated DNA, in which 100% of CpG are methylated, based on DNA from Human cells or from Hela cells.
- fully Un-Methylated DNA, in which 100% of CpG are methylated, based on DNA from Human cells

3.7.2.3 Bisulfite sequencing analysis

Sequencing data were obtained in a *.ab1 file format from the 3130xl Analyzer. The Data analysis software, Sequencing Analysis 5.2 (was used to process the raw data of which the results were displayed as an electropherogram of the base pair sequence (each base is represented by a peak in a different colour).

The percentage of methylation at individual CpG sites was calculated by dividing the height of the peak of cytosine signal (i.e. protected from conversion by the methylation) by the height of both the peak of thymine (converted as not protected by methylation) added to the height of the cytosine peak signal, and multiply by 100 as in the formula below.

% of Methylation = [C Peak/ (C Peak + T Peak)] x 100

Of note: The completeness of the bisulfite conversion reaction can be confirmed by the absence of free cytosine peak (i.e., a cytosine that is not at CpG site) as all free cytosine should be converted to Thymine.

3.8 *TNF-α* promotor Bisulfite sequencing condition

The conditions for the *TNF-α* promotor region sequencing after Bisulfite conversion were optimised with another student (co-author) and presented in my manuscript (284). Briefly, CD4 T-cells were isolated (CD4+ T cell enrichment kit STEMCELL) with ~ 97% purity. DNA was extracted (QIAamp DNA blood mini kit), and bisulfite converted using the EZ DNA methylation-GoldTM kit. To amplify a small region in the promoter of the *TNF* gene, a polymerase chain reaction (PCR) was performed (containing HotStarTaq enzyme, dNTPs, buffer, QIAGEN), 500 nM of forward and reverse primers (F5'-GAGTGTGAGGGGTATTTTTGATGTT-3'), (R5'-CTCTCCCTCTTAACTAATCCTCTA CTATCC-3'), 1mM MgCl2 and 2 µL of converted DNA. PCR conditions were 1 cycle of denaturation (10 min at 95°C), followed by 40 cycles of amplification (94°C for 10s, 59°C for 20s and 72°C for 45s), and a final extension (72 °C for 10 min). The PCR product (2.5 µL) was added to 1 µL of ExoProStar enzyme (IllustraTM ExoProStarTM) and placed in the thermocycler for 1 cycle of 15 min at 37°C, followed by 15 min at 80°C. The PCR product was finally diluted with 3.5 µL of nuclease-free water before use in the sequencing step. BigDye Terminator (V3.1 Cycle sequencing kits,) was used for sequencing: 0.25 µL of "Ready Reaction Mix", 3.5 µl of ABI sequencing buffer and 0.16 µM primer (forward and reverse reaction done separately) were add to 1 µL of PCR product in 10 µL total volume. The mix was placed on a thermocycler for 28 cycles (96°C for 10 sec, 50°C for 5 sec, and 60°C for 4 min). The samples were then precipitated with ethanol and the DNA pellets dried for 1 minute at 95°C. Pellets were re-suspended in 20 µL of HiDi formamide and sequenced on the 3130xl Genetic Analyzer.

3.9 Development of qMSP (Quantitative methylation-specific PCR)

3.9.1 Method Principle of qMSP

qMSP is a fluorescence-based, real-time qPCR method to detect DNA methylation at a locus in genomic DNA (258, 259). This technique employed the capability to distinguish methylated cytosine from normal cytosine after bisulfite-conversion and the capability of highly sensitive quantification of a target region by qPCR (258, 259, 282, 285). The detection of methylation status relies on methylation-specific primers (and also the probe when using TaqMan-based detection). The concept is therefore to prime and amplify only when a target locus is either methylated or unmethylated. This imply that specific CpG itself(s) is/are usually targeted by the assay (rather than a whole region)(286). The methylated (or demethylated) status of the target locus can be determine using the Ct value from the specific reaction (target-Ct) and the quantitative part needs to be calculated as the relative methylation value (ΔCt) of that specific CpG to the implication of a reference gene that is not dependent on methylation (reference-Ct)(286, 287).

There are two types of fluorescent reporter molecules to detect qPCR products for this technique, the SYBR green and the TagMan chemistry.

SYBR green-based detection

The SYBR green method uses a double-stranded DNA binding fluorescence dye to monitor the amplification of the PCR product (included in the PCR reaction reagents) (288). As the PCR progresses, more PCR product is generated. The SYBR green dye binds to double-stranded DNA amplicon, which results in an increase of fluorescence intensity proportionally to the amount of PCR product (289, 290). The SYBR green method is relatively low costs and easy to design. However, the binding of SYBR green dye is non- specific (it binds to any double-stranded DNA in the reaction) which can lead to false-positive fluorescence detection notably if primer dimers or other type of double strand DNA are present. Thus, well-designed primer and a post-PCR analysis for the amplicon's specificity is very important. This can be achieved using melting curve analysis and Agarose gel electrophoresis after the reaction to ensure the reaction specificity (i.e. the presence of a unique amplicon and the absence of primer dimers or heteroduplex)(291).

Melting curve analysis (Figure 3-4)

After a SYBR green PCR reaction, a melting curve analysis can determine the PCR specificity by observing how the fluorescence signals change while increasing the temperature. As the temperature raise, double-strand DNA which has incorporated the fluorescence dye molecules, dissociates into single-strand DNA and the dye molecules is released (hence not able to fluoresce any longer). Sudden decrease of fluorescence signal are detected when the melting temperature (Tm) of dsDNA amplicon is reached. Because this Tm is directly related to the length and GC content of the amplicon, the melting characteristics can be used to distinguished different sizes of DNA products (i.e. the amplicon versus primer dimers). Primer dimers are shorter than the expected PCR product thus have a lower Tm. Plotting the change in fluorescence / change in temperature (–ΔF/ΔT) against temperature provides a good view of melting dynamics. The peaks show a sudden change in signal when a Tm is reached. A single peak indicates a single product while multiple peaks indicate more than one product (288, 291).

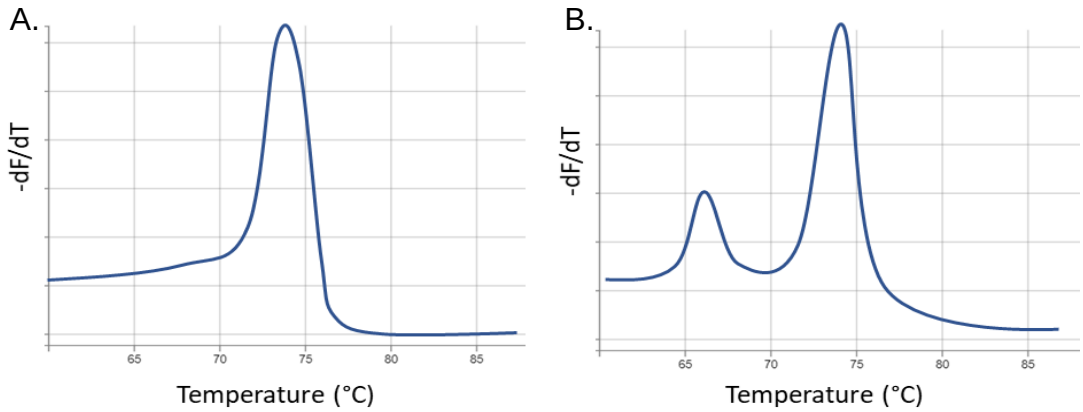


Figure 3-4 Melting curve analysis. A) A single peak indicative of a single PCR product. B) Multiple peaks indicative of more than one PCR product being produced.

Tag-Man based detection

The Tag-Man method (292) was designed to improve specificity of the Real-time PCR by introducing an oligonucleotide probe in the reaction. This probe is designed to the bind DNA sequence between the two primers (in the amplicon) and contains a fluorescent reporter dye on the 5' end of the probe and a quencher dye on the 3' end. While the probe is intact, the proximity of the quencher dye prevents the fluorescence emitted by the reporter dye through fluorescence resonance energy transfer. During the PCR reaction, the probe binds to the target sequence (single stranded amplicon). When the Taq polymerase synthesising the new DNA strand from the forward primer reaches the probe, it cleaves the probe using its intrinsic 5’ nuclease activity. The destruction of the probe causes the separation of the reporter dye from the quencher, allowing the reporter to produce a fluorescent signal. Each time a new PCR amplicon is created, the reporter and quencher are separated, and the fluorescence increases proportionally (288).

Fluorescence, therefore, increase proportionally with the product allowing effectively monitor of the reaction throughout the run (288). TagMan provides more specific detection and no need for post PCR check-up however it comes with the high cost and needs more effort for assay design.

qPCR analysis (Figure 3-5**)**

PCR theoretically doubles the number of amplicons with each amplification cycle. The fluorescence can be detected and recorded by the instrument at every cycle as the reaction progresses. The instrument then generates amplification curves by plotting fluorescence against the cycle number showing the accumulation of PCR products over the duration of the entire reaction.

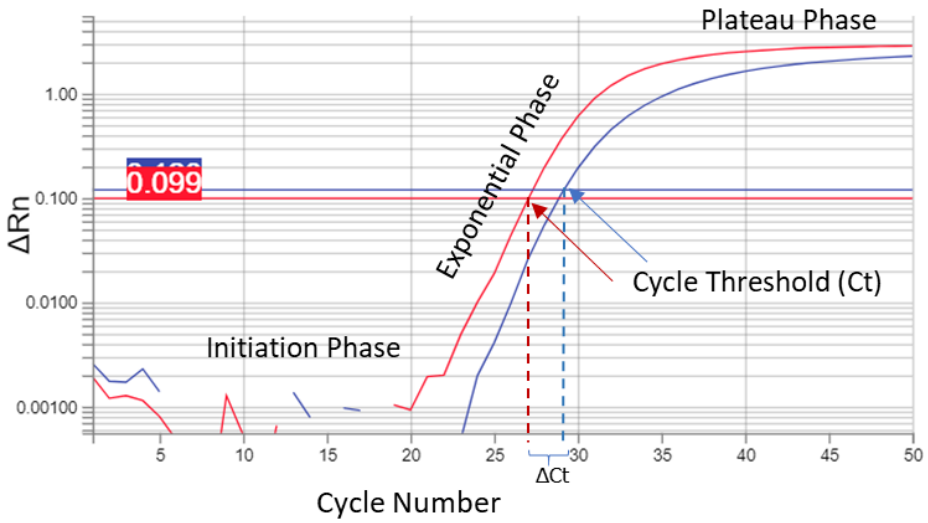


Figure 3-5 qPCR amplification plot of the target gene (blue) and reference gene (red) assay. qPCR amplification plot demonstrates the change in fluorescence over the number of cycles. It shows in sigmoidal curve compose of 3 phases. The cycle threshold (Ct) value of each reaction is defined as the cycle number when the fluorescence of a PCR product can be detected above the background signal in the exponential phase. The relative amount of PCR product of the target gene can be measured by comparing Ct of target gene reaction to the reference gene reaction (ΔCt).

The PCR amplification curve develops over 3 phases :

- initially a background phase at the beginning of the reaction where there is no fluorescence as the amount is below detection capacity of the instrument,
- an exponential phase where the amplification is at full efficiency, and fluorescence levels double at each cycle to the next,
- a plateau phase where the reaction is limited by the exhaustion of reagents.

Data collected during the exponential phase of the reaction allows the user to determine the initial quantity of the amplification target precisely.

A reaction threshold needs to be set in the exponential phase where the level of fluorescence is directly proportional to the cycles and where the PCR efficient is theoretical 100% efficient. The quantity of an amplification determined by a certain amount of florescent is chosen (a level significantly away from the background phase) that is related to a specific cycle of the PCR reaction : a cycles threshold or Ct, which is the cycle number at which the chosen fluorescent levels is achieved. Ct values are directly related to the starting amount of target sequence. The higher the Ct value, means that more cycles are needed to amplified the amount of PCR product needed to reach the fluorescence level chosen.

Relative amount of PCR product are usually used to described data. This is done by using a reference gene to be amplified under the same condition but for which the initial amount is only related to the amount of input material. It is used to normalize samples. For gene expression this is usually performed using housekeeping gene(s), which expression is not regulated by any factor and therefore is the same in every cells. In the case of qMSP it should be a region of DNA that does not include any CpG hence not susceptible to change in methylation status.

qMSP Reaction

qMSP reactions are an adaptation of the qPCR Principe. Both SYBR green and TaqMan chemistries can be used. Two types of reactions are needed for a quantitative assay: one assay, methylation-dependent for a CpG of interest and a methylation-independent assay for a control gene for normalization.

The detection of the methylation status of the target genes by qMSP relies on methylation-specific primers that are designed to prime and amplify only when the target CpG is either methylated or unmethylated. The assay is developed by designing primers covering a region with several CpG sites. The more CpG sites (with the same methylation status), the more specific the assay. This applied to both design for a SYBR green or TaqMan assay.

To achieve accurate and reliable qPCR results when working with the patient DNA samples which vary in initial template quantity and integrity, especially after bisulfite treatment, normalization of the target CpG against a control assay (for a region not susceptible to change in DNA methylation) is critical to adjust for these

difference in quantity and quality of DNA input. The Internal control for qMSP could be any sequence of DNA that is not containing CpGs. The control region assay can be performed in a separate reaction parallel to the reactions of the target gene for an individual patients sample.

Quality control and assay development

Optimizing qMSP reaction as well as working on multiple samples (multiple plates) requires the use of standard DNA sample with known levels of methylation that is achieved using completely methylated or completely unmethylated DNA derived from cells (or cell lines) and commercially available. During optimization it is needed to see the specificity of methylation-dependent primers and for creating a dilution curve to access the assay efficiency. No template control is also used in the assay development and to check contamination or any false positive signal (particularly when using SYBR green to detect primer dimers).

Efficiency of the PCR assay

In an Ideal PCR reaction, the amount of template should double in every cycle during exponential amplification. The efficiency, which is an ability to double the amount of template in every cycle, should be 100%. However, in the practical PCR reaction, the experimental factors such as the secondary structure, non-optimal reagent, PCR inhibitor can influence the reaction efficiency (293). Therefore, the efficiency of the newly developed reaction should be tested before the assay can be used.

PCR efficiency of each assay could be obtained from the dilution curve which is the plot between the log of the serial of DNA concentration and its Ct value. Fitting the standard curve to linear regression model provides information on the slope, y-intercept and correlation coefficient (R^2) of the curve. The reaction efficiency was calculated from the slope using the formula below.

$$\text{Efficiency} = 10^{(-1/\text{slope})} - 1$$

Efficiency varies between reactions and primers pairs / concentration. The reactions should have efficiency as close to 100% (a slope of –3.32) as possible. Efficiencies between 90% and 110% were considered a good reaction and then was chosen to continue to work with the real clinical sample (288).

The y-intercept corresponds to the theoretical limit of detection of the reaction. It may be useful for comparing different amplification systems and targets. R^2 could provide how well the data fit the regression model. The closer to 1 the better. In some case, data at the edge of dilution series (either at high or low concentration) could be removed from the plot to improve the R^2. The range of DNA concentration in which R^2 still in a good fit implies the sensitivity of that PCR assay. It provides the dynamic range of DNA template that can be used in PCR reaction where it does not affect the efficiency.

3.9.2 Method detail of qMSP

qMSP is used to quantify the methylation status in region containing a candidate CpG site. The Development of qMSP started with assay design, reactions optimisation and efficiency test which were done using fully unmethylated and methylated control DNA. The Optimized assay was then used to quantify methylation of the target locus in genomic DNA from patients samples. The design and optimisation were performed for two qPCR detection method, SYBR green and Tag-Man, as separately described below.

3.9.2.1 Design and optimisation of a SYBgreen assay

3.9.2.1.1 Assay set up

For the individual samples, two types of reaction are required in order to determine the methylation of the target genes using qMSP assay.

- a PCR reaction for the gene of interest which is a methylation-dependent reaction designed to interrogate the methylation status of a specific CpG.
- a PCR reaction for an internal control which is a methylation-independent reaction designed to normalise the target gene reaction to the input DNA in the sample.

3.9.2.1.2 Primer design

qMSP primer was designed following the basic rules for amplifying bisulfite-treated DNA. This includes increasing the reaction specificity by increasing primer length or introducing more Guanine in the primer sequence), decreasing the PCR product size, and/or increasing cycle number (294).

For the PCR reaction of the target genes, qMSP primers were aimed for detecting **the <u>de</u>methylated DNA** at the specific CpG site targeted. Thus, the primers were designed to recognize the unmethylated CpG site targeted and located at 3' end of the primer. Because unmethylated cytosines were converted to Thymine (after DNA bisulfite conversion), the primers sequence for the CpG site were then designed to detect TG not CG.

For the PCR reaction of the internal control, the qMSP primers should amplify bisulfite modified DNA independently its methylation status. They were therefore designed to avoid regions containing CpG sites. Example of primers design on target gene and control genes shows in Table 3-4.

To do this, First, the genomic DNA sequence of candidate genes was obtained from NCBI database. The sequence was then bisulfite converted *in silico* using the sequence manipulating function in MethPrimer 2.0. Forward and Reverse primer were designed using the, Primer Express 3.0. The primers sequences were manually modified to obtain the best Tm difference, %GC, position of the CpG site in the primer, the number of free cytosines, and most importantly to ensure no primer dimers/loop or secondary structure.

The *In silico* PCR product was checked to ensure specificity as the absence of homology present with other sequences on the bisulfite-converted human genome, using the blasting function in Bisearch primer-design and search tool.

All primers were synthesised by Thermo Fisher Scientific, reconstituted to a concentration of 100 uM in sterile water. Primers were aliquoted stored at -20^{o}C. The detail of each primer pair was described in the result section and Appendix 5.

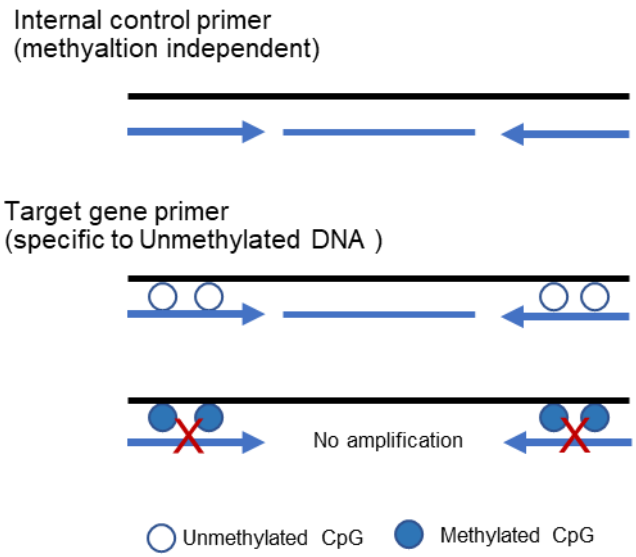


Table 3-4 Example of SYBR green qPCR primers design on target gene and control genes.

3.9.2.1.3 General SYBgreen-based qMSP reaction

After bisulfite conversion, genomic DNA was amplified using locus-specific PCR primers. Power SYBR™ Green PCR Master Mix was used. The standard PCR mixture and Thermo cycle condition recommend by the manufacture's protocol were followed as described in Table 3-5 and Figure 3-6. The PCR reaction was performed in MicroAmp™ Optical 96-Well Reaction and run on QuantStudio 5 Real-Time PCR Systems. To confirm the reaction specificity, melting curve analysis was also performed on the same machine right after the PCR reaction. This informs on the presence of a single PCR amplicon (hence specificity) as well as presence of primer dimers even if the design was meant to limit them. The data was collected and analyses by " The Design and Analysis II " module. The threshold cycles (Ct) method was used to quantify the product and normalise it to the control PCR. After the thermal cycling reaction, Gel electrophoresis was also performed to validate the data obtained from a melting curve.

Table 3-5 Reaction composition and qPCR cycling program of SYB Green-based qMSP

A: Composition of qMSP reaction

Stock conc.	Reagent	Final conc.	Vol/reaction (ul)
2X	SYBR mastermix	1X	12.5
vary	F primer	50-900 nM	2.5
vary	R primer	50-900 nM	2.5
10 ng/ul	DNA template	50 ng/ul	5
	Distilled water		2.5
	Total Volumn 25 ul		25

B: qPCR cycling program

qPCR cycle

Step	Temperature	Time	
Initial activation	95 ºC	10 min	
Denaturation	95 ºC	15 s	40 cycles
Annealing/Extension	60 ºC	60 s	
Melt curve	95 ºC	15 s	
	60 ºC	60 s	
	95 ºC	1 s	

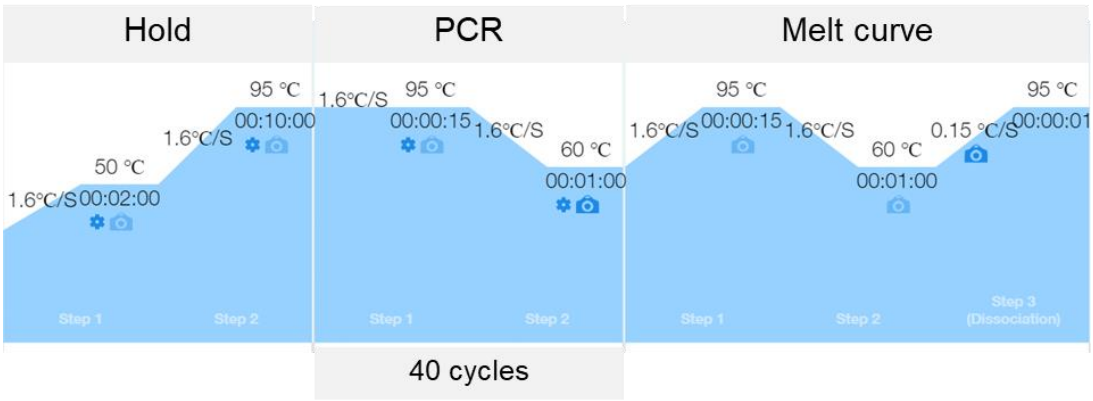


Figure 3-6 Diagram of SYBgreen-based qMSP cycling program

3.9.2.1.4 Optimisation

For each SYBR green based qMSP reaction, several factors need to be optimised to obtain a efficient assay. The fully unmethylated DNA control was used during this optimisation process.

- Primer concentration: The purpose of this procedure is to determine the minimum primer concentrations giving the maximum ΔRn.
    - F and R primer were tested between 50 mM-900 nM according to the primer matrix below.

Forward Primer

Reverse Primer	**50 nM**	**300 nM**	**900 nM**
50 nM			
300 nM			
900 nM			

- Annealing temperature is optimal in SYBR green qPCR assay at 59-60^{o}C. Adjusting annealing temperature of the primers was performed to increase the reaction specificity in all primer set.
    - The reaction was run at different annealing temperature ±3^{o}C from the standard annealing temperature at 60^{o}C.

It is important for SYBR green reaction to ensure of the absence of primer dimer or any non-specific product that can lead to the false positive signal.

3.9.2.1.5 Efficiency and specificity of the PCR assay for un-methylated DNA

PCR efficiency of both the target gene and internal control reactions were assessed by creating a dilution curve. A serial dilution of fully methylated and unmethylated DNA controls was performed at a concentration range from 0.01 ng to 50 ng. The reaction was performed in triplicate for each condition. Ct value of each points in the dilution series was collected and plotted against the log of DNA concentration to generate a standard curve for both the target and reference genes. The assay efficiency of both assays was calculated as mentioned earlier.

PCR specificity for the target gene and the control gene could also be observed from this experiment. The internal control reaction should be able to amplify both methylated and unmethylated DNA sample with the same efficiency, while the

target gene reaction should only be able to amplify specifically the unmethylated DNA.

3.9.2.2 Design and optimisation of Tag-Man assay

3.9.2.2.1 qMSP Assay set up

Two types of reaction are again required in order to determine the methylation of the target genes using qMSP assay.

- a PCR reaction for the gene of interest which is a methylation-dependent reaction designed to interrogate the methylation status of the targeted CpG.
- a PCR reaction for the control gene which is a methylation-independent reaction designed to normalise the target gene reaction to the input DNA in the sample.

3.9.2.2.2 Primer/probe design

qMSP primers were designed following the same basic rules.

However, for this TaqMan PCR reaction for the target CpG, qMSP primers and probe were aimed for detecting and amplifying **the methylated DNA** at specific CpG sites (mainly due to need for high melting temperature difficult to achieve on the unmethylated sequence). Because of methylated cytosine will still remain as a cytosine after bisulfite conversion, the primers should be designed include 3-5 CG sites and should detect CG (not TG as in the previous SYBR green assay that was detecting the un-methylated DNA). Location of the CpG site at 3'end of primer help increases specificity. Probe also can be designed to include CG site to increase the reaction specificity. For the PCR reaction of internal control, qMSP primers should amplify bisulfite modified DNA independent of methylation. They were therefore designed to avoid CpG rich regions.

The genomic DNA sequence surrounding the candidate CpG was obtained and converted as previously described. Forward primer, Reverse primer and probe were designed using the TagMan MGB quantification function of the primer design software, Primer Express 3.0. The primers/probe sequences were manually modified to obtain the better Tm difference, %GC, position of CpG site, and the number of free cytosines. The *In silico* PCR product was checked to ensure the specificity to the converted genome as before. All primers were synthesised by Thermo Fisher Scientific, reconstituted to a concentration of 100

uM in sterile water. TagMan custom probe-MGBNFQ were synthesised by Applied Biosystems reconstituted to a concentration of 2.5 uM in sterile water. Primers and probe were aliquoted stored at -20^{o}C. The detail of each primer pair and probe design, are described in the relevant results sections

3.9.2.2.3 Tag-Man PCR reaction

After bisulfite conversion, genomic DNA was amplified using locus-specific PCR primers and TaqMan™ MGB Probe. TaqMan™ Universal Master Mix II, (no UNG version) was used. The standard PCR mixture and Thermo cycle condition recommend by the manufacture's protocol are described in Table 3-6 and Figure 3-7. The PCR reaction was performed as before. The data was collected and analysed as before.

Table 3-6 Reaction composition and qPCR cycling program of TaqMan-based qMSP

A: Composition of qMSP reaction

Stock conc.	Reagent	Final conc.	Vol/reaction (ul)
2X	universal mastermixII no UNG	1X	10
vary	F primer	50-900 nM	2
vary	R primer	50-900 nM	2
2.5 uM	Taqman probe	250 nM	2
	DNA template	10 ng/uL	2
	Distilled water		2
	Total Volumn 20 ul		20 ul

B: qPCR cycling program

Step	Temperature	Time	
Initial activation	95 ºC	10 min	
Denaturation	95 ºC	15 s	45 cycles
Annealing/Extension	60 ºC	60 s	

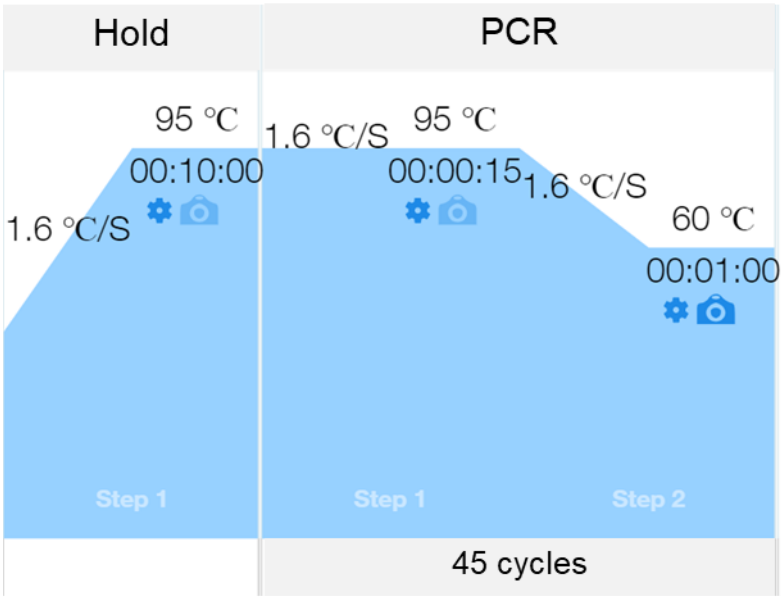


Figure 3-7 Diagram of TaqMan-based qMSP cycling program.

3.9.2.2.4 Optimisation

For each TagMan reaction, several factors needed to be optimised the same was as the SYBR green reaction. The fully methylated DNA control was however, used during the optimisation process.

- Primer concentration: tested between 50 mM-900 nM.
- Annealing temperature: tested ±3^{o}C from the standard 60^{o}C.

3.9.2.2.5 efficiency of the PCR and specificity for methylated DNA

PCR efficiency of TagMan assay was performed the same way of for the SYBR green-based qMSP as described earlier.

PCR specificity checking for the control gene reaction in TagMan assay was also the same with SYBR green-based assay. However, the TagMan target gene reaction should be able to amplify specifically to methylated DNA instead of unmethylated DNA.

3.9.3 Assay performed on patients samples

Once I had optimised the final condition for the assay, it was used to access methylation of the target CpG in genomic DNA.

For each sample, target gene reactions were run in parallel to the control reaction and were performed in duplicate/triplicate. Fully methylated and unmethylated control DNA were included on each plate as a positive and negative control. A no-template control was also added to ensure no contamination.

3.9.4 qMSP quantification

The relative level of methylation in the target gene could be measured using Ct value of the reactions. It is presented as a percentage of methylation (% M). The formula to calculate it is as followed.

Percentage of methylation (%M) = Relative level of methylation x 100

= $2^{-\Delta\Delta Ct}$ x 100

By;

ΔCt sample = Ct_{sample} target gene - Ct_{sample} internal control

ΔCt calibrator = $Ct_{calibrator}$ target gene - $Ct_{calibrator}$ internal control

$\Delta\Delta Ct$ = ΔCt sample - ΔCt calibrator

Relative level of methylation = $2^{-\Delta\Delta Ct}$

Sample	is an individual bisulfite converted DNA sample.
Calibrator	is the standard fully methylated bisulfite converted DNA refer as 100% methylation
Target gene	is the methylation-dependent reactions of each target genes
Internal control	is the methylation-independent reactions of the internal control to normalized target gene reaction with the input DNA of the individual sample.

3.10 Statistics

Highly specific statistics were applied all the way through the thesis and were described in the relevant sections. This notably applies to the bioinformatics analysis performed in my thesis. The development of a Biomarker requires specific tools as well which are described here in details.

Different statistic tests were used for comparing the % of methylation between different patients groups according to each analysis. Non-parametric tests were used, Mann-Whitney U (MWU)(unpaired), Wilcoxon signed-rank test (paired), and Dunn test (≥3 groups) followed by Kruskal–Wallis test for multiple corrections were used as data were not normal distributed. The statistical significance was obtained when the p values are 0.05 or less. Statistical analysis was performed using R software 3.5.2.

Statistic involved in evaluation of Biomarker performance

A biomarker can predict whether or not that patient has a particular outcome (a disease, or a risk for something or a particular response to treatment). Several statistical characteristics are used to evaluate biomarker performance. Receiver Operator curve (ROC) and area under the curve are used to access the predictive or discrimination ability of biomarker (295). Performance characteristics such as sensitivity and specificity describe how well the test is able to identify patients with the target outcome, while positive and negative predicted value gives information about the value of a specific test (296). These scores are calculated by comparing the biomarker predicted result to the actual outcome (as further detailed below).

In order to obtain a predicting value for a biomarker, a model of the interaction between the outcome (or response variable) and the biomarker (predictor variable) have to be developed. The statistical tools used to help with this analysis are usually logistic regressions.

Logistic regression is used to examine the relationship between a dependent categorical variable (the outcome) and one (or more) independent variable(s). It is a method of choice to evaluate biomarkers in clinical research when looking at a binary outcome (297) e.g. a biomarker used to determine whether or not (“yes” or “no”) patients are having the outcome (a disease, or a “ good response” or “poor response” to an intervention). This kind of analysis is known as a binary logistic regression.

Logistic regression fit models for the probability of an event occurring (the clinical outcome) depending on the values of the independent variables (the biomarker and other variables). Logistic regression also attempted to

1) estimate the probability (P) that an event occurs vs does not occur,

2) predict the effect of an independent variable(s) on a binary outcome (increase or reduce the risk),

3) classify observations in a particular category or another (i.e. high/low risk).

Odds ratios (OR) obtained from logistic regression can be used to describe the strength of associations of biomarkers with clinical events. OR describe how odds of having a particular outcome change with an increase of 1-unit in the biomarker measurement (297, 298). OR = 1 means there is no association between the marker and the clinical outcome. OR > 1 means greater odds of association between the marker and the clinical outcome, while OR < 1 indicate an association in the opposite direction (i.e., protective against the outcome).

Logistic regression model provides the probability (P) of the event at a giving value of the overall sets of biomarker/variable included in the model. To provide the predicting performance result of the model, a probability threshold needs to be set. When the probability is greater than the set threshold, the event is predicted to happen otherwise it is predicted not to happen.

Confusion matrix or classification matrix

The prediction ability or classification performance of the logistic regression model can be described using a confusion matrix which is a table comparing the actual outcomes to the model (predicted outcomes) with the actual outcome itself (observed outcome)(295) Figure 3-8. At a selecting threshold, the model predicts positive or negative events. After comparing these number to the actual outcomes, the number of True positives (TP), True negatives (TN), False positives (FP), and False negatives (FN) events are obtained (confusion matrix) and can be used to calculate the performance of the binary classification test as below and in Figure 3-8;

- **Sensitivity**, also referred to as true positive rate, measures the proportion of actual positives that are correctly identified.
- **Specificity**, also referred to as true negative rate, measures the proportion of actual negatives that are correctly identified.

- **Accuracy** measures the total number of predictions that are correct.
- **Positive predictive value (PPV**) is the probability that the disease is present given a positive test result.
- **Negative predictive value (NPV)** is the probability that the disease is absent given a negative test result.

Sensitivity and Specificity are used as criteria for biomarker validation in the development process, however, for the clinician who has a test result in hand, PPV and NPV are more useful.

Difference threshold setting will affect the model result. The threshold should be selected according to the use of the test or the specific scenario where more false negative or false positive may be more or less acceptable.

ROC

The overall performance of a particular biomarker (or that of a model) can be evaluated by summarizing the results created by different sensitivity/specify thresholds in one curve. This curve, called receiver operating characteristic curve (ROC, Figure 3-9), plots True Positive Rate (or sensitivity) versus False positive rate (or 1- specificity) at all possible threshold. This curve shows the overall tread off between specificity and specificity (295) and can be used for determining the best cut off value for predicting whenever a new observation is a YES or a NO.

The area under the ROC curve (AUC or AUROC) are used to evaluate the overall prediction or classification performance of a logistic regression model (297). The AUROC value can range from 0 to 1. An area close to 1 (comprised between 0.5-1) indicates a perfectly correct prediction *of* the outcome of the test/model and increasing performance in classification getting closer to 1 and while an area of 0.5 indicated a random prediction. An area tending to 0 (comprised between 0 - 0.5) indicates a perfectly correct prediction *against* the outcome (i.e. a protective value) and best performance in classification tending to 0.

AUROC notably helps investigators compare between different logistic models to decide which (or which biomarkers) is best or to determine whether the new biomarker has an added value compared to the exist biomarkers or risk factors.

Actual outcome

Predicted outcome	Positive	Negative	
Positive	True Positive (TP)	False Positive (FP)	**Positive predictive value (PPV)** $\frac{TP}{(TP+FP)}$
Negative	False Negative (FN)	True Negative (TN)	**Negative predictive value (NPV)** $\frac{TN}{(TN+FN)}$
	Sensitivity $\frac{TP}{(TP+FN)}$	**Specificity** $\frac{TN}{(TN+FP)}$	**Accuracy** $\frac{TP+TN}{(TP+TN+FP+FN)}$

Figure 3-8 Confusion matrix and the derivation of main diagnostic parameters

Figure 3-9 ROC curve. Comparing ROC curve with different classification and prediction performance.

Chapter 4 Results Part1 : Genome-wide DNA methylation data analysis

4.1 Introduction

To gain more understanding of whether DNA methylation is involved early in the RA pathology, I analysed DNA methylation patterns of ~480,000 CpGs, in 3 cell types (naive CD4+T-cells, memory CD4+T-cells, and monocytes) from 6 HC and 10 early RA patients obtained from an Illumina methylation genome-wide array.

In this chapter, I will describe the profiles of DNA methylation in HC and RA in 3 cell subsets and the strategy that I developed to find differentially methylated CpG (DM-CpG) as an exploratory phase. I will then describe the further downstream analysis and some experiments performed to validate the analytic results and to link DM observed in RA to disease pathogenesis. An overall analysis and experimental workflow is shown in Figure 4-1.

This dataset (48 genome-wide DNA methylation profiles) was obtained by my supervisor prior to my arrival in her group. I analysed it using the combination of standard analysis workflow for this platform and in-house R scripts.

I would like to make a point at the beginning of this chapter that parts of this work have been published in a manuscript entitled “Differential CpG DNA methylation in peripheral naïve CD4+ T-cells in early rheumatoid arthritis patients” on 07 April 2020 in Clinical Epigenetics journal. The manuscript also included some works done in collaboration with others members of the group and data from my supervisor’s previous work. The content included in my Thesis represents the work I have performed and the data obtained in collaboration will be mention where appropriated in the chapter’s discussion or otherwise in the appendix.

Demographic and clinical data details of the HC and RA samples used to sort the 3 cell subsets analysed by the 450 K DNA methylation genome-wide array are described in Table 4-1. This is typical of an early RA presentation although my supervisor took particular care in selecting patients with at least 3 swollen joints and raised inflammatory makers (CRP>10 mg/L) hence a slightly long disease duration (13 months).

Table 4-1 Demographic and clinical data for the control and RA patients used in the DNA methylation bead array

Cohort-1 :	**HC (n=6)**	**RA (n=10)**
age (years)*	42 (38-47)	50 (40-74)
M/F	3/3	7/3
ACPA (Pos/Neg)	na	6/4
Duration (months)*	na	13 (3-24)
TJC	na	10 (3-16)
SJC	na	3 (3-11)
CRP (mg/L)	na	20 (10-40)

Data are presented as the median (range)

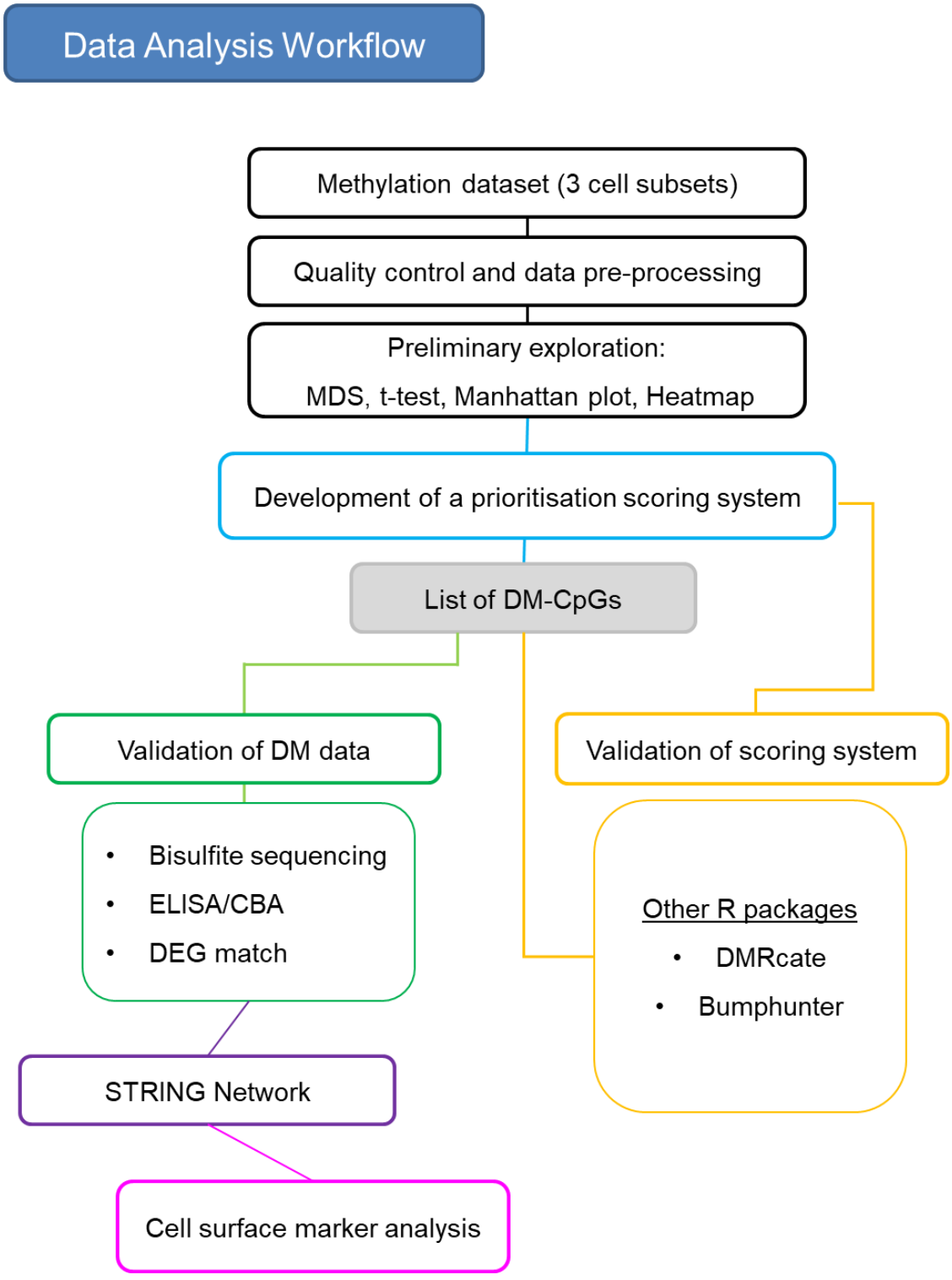


Figure 4-1 DNA methylation data analysis workflow.

4.2 Objective

For the first part of my PhD, my overall aim is to gain more understanding of early events/pathways involved in disease pathology by studying genome wide data on DMA methylation.

The **hypothesis** behind this is that alterations in DNA Methylation are central to the early events leading to RA pathogenesis rather than a consequence of its development and may therefore drive the development of chronicity. Secondly these alterations will affect CD4+T-cells, and preferentially naïve cells, considering on one hand that T-cells genes are the mainly targeted by the genetic susceptibility associated with RA and on the other, that naïve cells are more epigenetically uncommitted Th0 cells. Inflammation represents the hypothetical epigenetic injury trigger, leading to these epigenetic alterations.

Objectives are to:

- Explore genome-wide DNA methylation data in naïve and memory CD4+T-cell as well as monocytes in early RA compared to HC
    - Establish list of DM CpGs in the 3 cells types
- Validate DM CpGs
    - DNA level (using bisulfite sequencing)
    - RNA level (compared with publicly available gene expression data)
    - Protein level (using Elisa)
- Validate DM selection strategy using publicly available tools (R packages)
    - Obtain final list of DM CpGs aligned to genes symbols
- Develop a strategy to apprehend methylation changes with a biological meaning towards understanding better the early RA pathogenesis.
    - Establish a list of pathways targeted by DM
    - Explore *in silico* functional interactions between DM gene products.
- Additional work for understanding the biological meaning of DM (using flow-cytometry) is also included in the discussion of this part-1 of my thesis that has been performed in collaboration with other students (included on publication).

4.3 Result

4.3.1 Preliminary exploration of DNA methylation data

4.3.1.1 Quality control of the dataset

After constructing β-value histograms and bean plots, two samples (naïve CD4+T-cells of HC group) showed an abnormal distribution of the methylation level (β-value), (Figure 4-2) and had to be removed from the datasets.

On each DNA methylation profile, the probes which were related to common SNP and/or known to have cross-reaction effect were identified in the QC procedure and a total of 45,022 CpG sites were filtered out prior to any further analysis resulting in the final dataset of 440,490 CpG sites in 46 samples (4 HC and 10 RA of CD4+ naïve T-cell samples, 6 HC and 10 RA of CD4+ memory T-cell samples, and 6 HC and 10 RA of CD14+ monocyte samples).

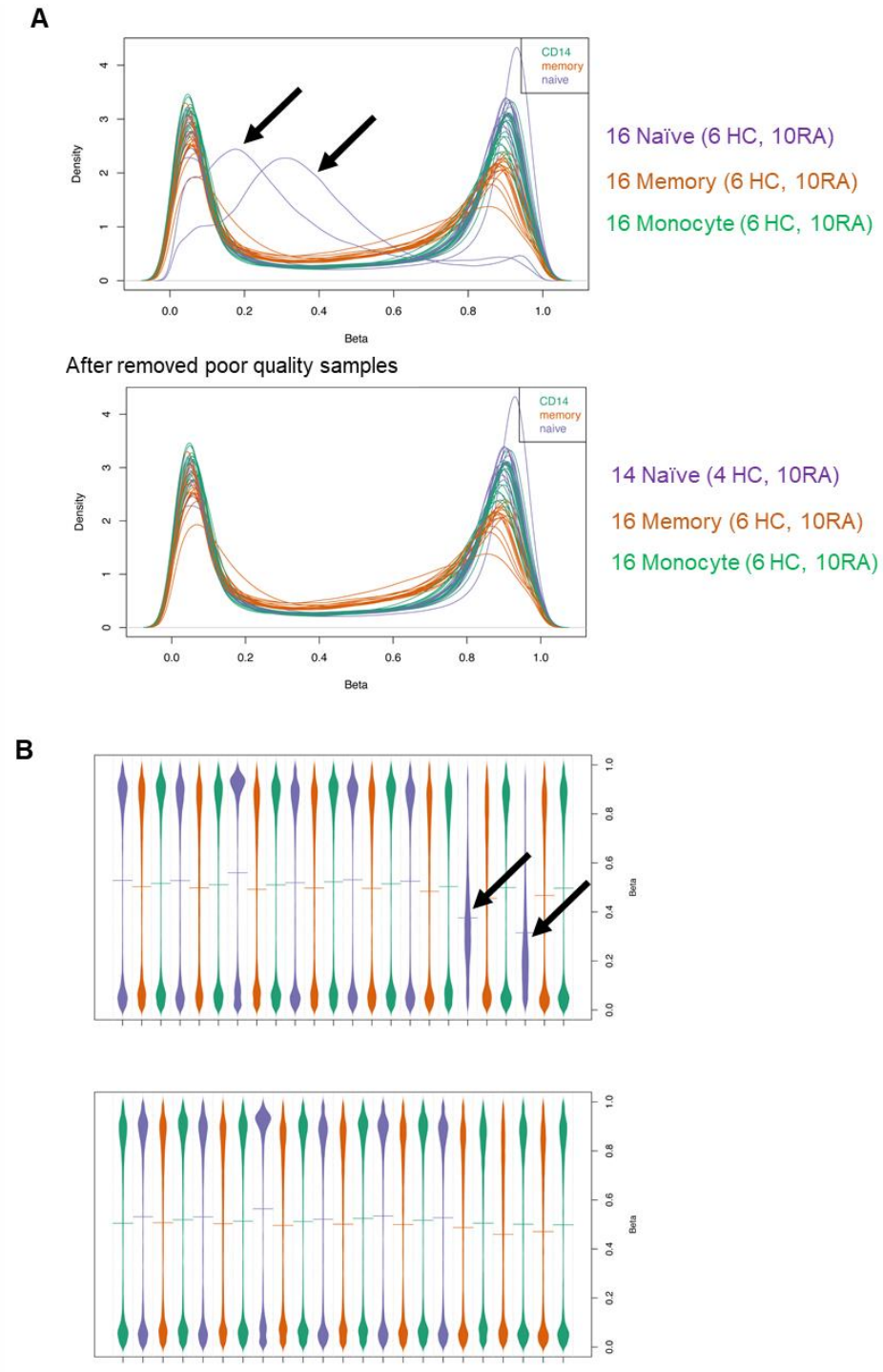


Figure 4-2 Quality control and data pre-processing A) β-value histograms and B) Bean plot for the 48 samples (naïve CD4+ T-cell (6HC,10RA), memory CD4+ T-cell (6HC, 10RA), and CD14+ monocyte (6HC, 10RA)). DNA methylation value should be distributed between only 2 values; 0-unmethylated or 1-methylated. The 2 samples (naïve CD4+T-cells of HC group) with intermediate density values (black arrows) failed the quality analysis and were excluded.

4.3.1.2 Preliminary exploration

The datasets were initially explored by multi-dimensional scaling (MDS) plot. The samples clustered tightly by cell type separating monocytes from naïve and memory T-cells (Figure 4-3,A). MDS did not segregate samples with respect of any gender bias (Figure 4-3,B, after exclusion of Chromosome X and Y data) or for disease groups (HC/RA) (Figure 4-3,C), although, there was some degree of segregation by PCA between HC and RA in monocytes that was not observed in T-cell subsets.

Methylation patterns are therefore more specific to the cell type than it is to other factors such as gender or disease. Interestingly, the PCA/MSD analysis resulted in a large separation between the 2 types of CD4+T-cell (naïve and memory) away from the monocytes. All further analysis to determine differences in methylation between HC/RA will be performed in the individual cell subsets.

The methylation difference between disease and HC for each CpG was examined using t-tests, in the individual cell subset datasets. This generated p-values for each CpG. In every test, there is always a chance that a result indicates a difference between two groups while no real difference actually exists (false positive, or type I error). In statistical analyses where a large numbers of tests are performed, the number of false positives increase and controlling of this number is highly recommended. However, correcting for multiple testing using the false-discovery-rate (FDR) method was too conservative in my study, and led to a very low number of significant In order to obtain a predicting value for a biomarker, a model of the interaction between the outcome (or response variable) and the biomarker (predictor variable) have to be developed. The statistical tools used to help with this analysis are usually logistic regressions. Therefore in the exploration phase, I decided not to apply the FDR and non-adjusted p-values were used.

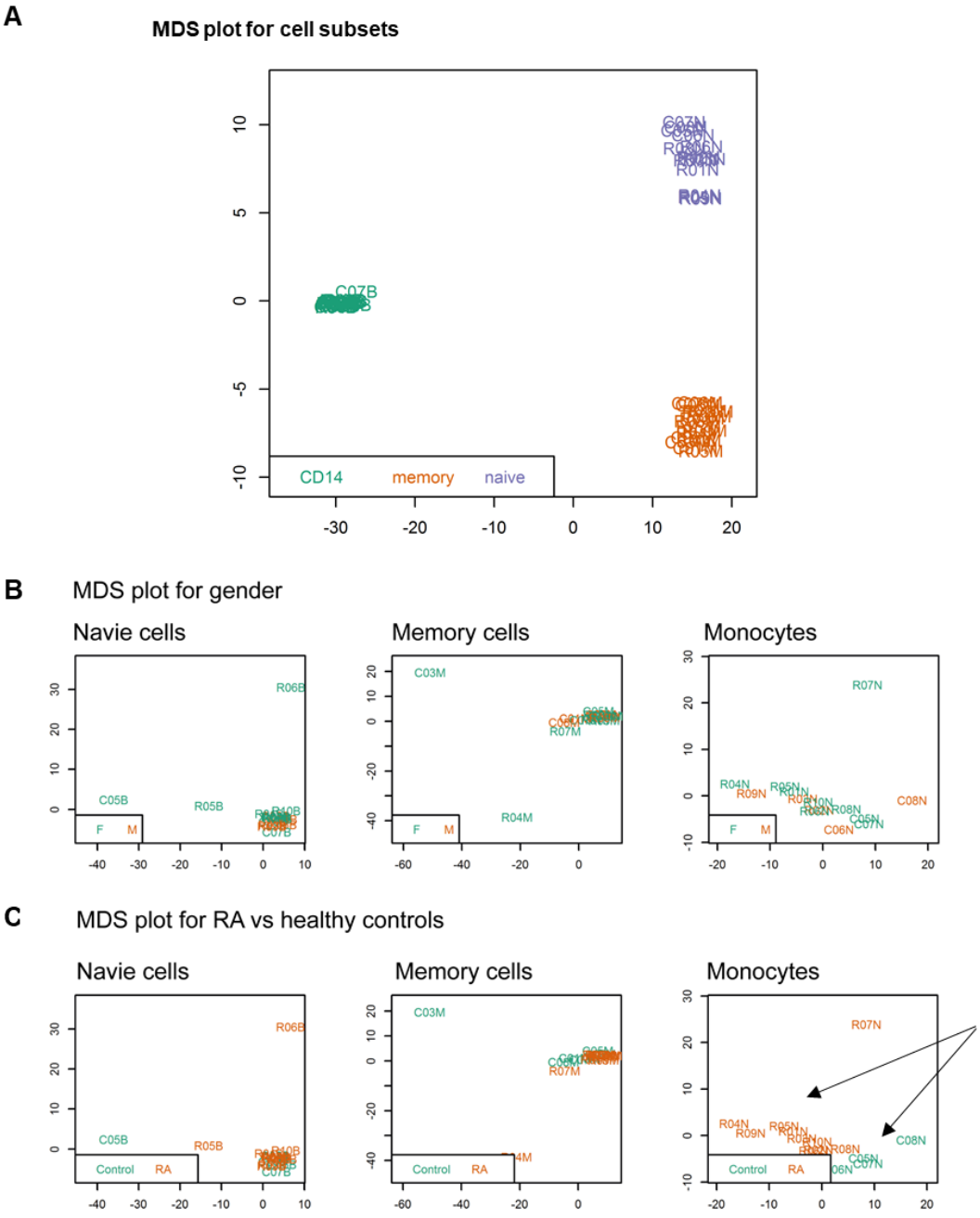


Figure 4-3 Preliminary exploration of datasets using MDS for A) cell subset, B) gender and C) RA versus HC. Data were clearly segregated by cell types. Analysis by gender (in individual cell types) showed no discrimination similarly to analysis by disease for T-cells although monocytes showed some degree of separation (arrows).

The levels of DM that could have a significant effect were then explored in an ordered manner on each chromosome. The overall methylation differences between HC and RA across the genome were explored using Manhattan plots in the 3 individual cell types. The plots highlighted clouds of data points departing from the axis (high values for –log10(p-value) above a region with very high density of dots (low values for –log10(p-value). Three thresholds were set for different levels of the p-value; highly significant ($p \leq 0.0001$), medium significant ($p \leq 0.001$) and low significant ($p < 0.01$). The p-value indicates the probability of detecting a false-positive. Setting the cut-off p-value at a lower level (highly significant) help obtaining results with lower chance of being a false positive, thus more reliable results. In this dataset where the total test number of each cell subset is 440,490 CpGs, the false-positive rate from using a highly significant threshold would be 44 CpG while at the medium and low significant would present at 440 and 4,404 CpGs, respectively. In the Manhattan plot, there was a number of data-points (Table 4-2 561 data point in naïve cells) distributed at a higher-magnitude values for highly significance set at above -4 (Figure 4-4. corresponding to $p \leq 0.0001$), a larger number of data-points with medium significance (2,891, $0.0001 < p \leq 0.001$), and an even larger with low significance (14,568 , $0.001 \leq p < 0.01$) above the main region of highly dense data-points The number of CpGs passing each threshold were much more than the expected number of false positive suggesting that there are real differential methylations of the CpGs between HC and RA, which did not just happened by chance. This plot separates differentially methylated-CpGs (DM-CpGs) more clearly from the background. The number of DM-CpGs for the 3 cell types shows in Table 4-2.

In naïve T-cells, the number of DM-CpGs (18,020, $p \leq 0.01$) was the highest (4.1% of total tested CpG). Of this number 3.11%, 16.04% and 80.84% were categorised into high, medium, and low threshold of significance, respectively. After annotation, I analysed the distribution of DM-CpGs with respect to their location in gene structures for (i) core of CpG-islands (ii) shelves/shores of an island or (iii) in open sea (i.e. outside of a defined island). An equal distribution of DM-CpG in each category was observed. There is also similar proportion of hypo and hyper-methylation, with a slight bias toward hypomethylation.

In memory T-cells the main difference was that DM preferably occurred in the core of islands (50%) and were mainly hypermethylation (93% of all DM) which could possibly suggest more generalised gene silencing. In monocytes, DM was of lower significance altogether (Table 4-2).

A hierarchical clustering analysis displayed as heatmaps (Figure 4-5) showed a clear segregation of patients and HC for individual cell subset, as well as major

hypermethylation (green in HC versus red in RA) in memory T-cells. Overall, these explorations demonstrate distinct patterns of methylation changes between cell types and between HC and RA.

Table 4-2 Summary of differential methylation at individual CpG level

	Naïve CD4 T-cell	Memory CD4 T-cell	Monocytes
Number (%) of differentially methylated % total probes tested (440 490)	18 020 4.09 %	14 197 3.22 %	6 490 1.47 %
Number (%)* of probes with			
high significance (p ≤ 0.0001)	561 (3.11%)	440 (3.10%)	130 (2.00%)
medium significance (0.0001<p ≤ 0.001)	2 891 (16.04%)	1 885 (13.28%)	672 (10.35%)
low significance (0.001<p ≤ 0.01)	14 568 (80.84%)	11 872 (83.62%)	5 688 (87.64%)
Number and (%)*of probe associated with			
core island	6 141 (34.08%)	7 120 (50.15%)	1 985 (30.59%)
shelves/shore island	5 873 (32.59%)	3 948 (27.81%)	2 060 (31.74%)
outside of CpG island	6 006 (33.33%)	3 129 (22.04%)	2 445 (37.67%)
Hypermethylation in RA : n, (%)*	8 425 (46.75%)	13 218 (93.10%)	3 525(54.31%)
Hypomethylation in RA : n, (%)*	9 595 (53.25%)	979 (6.9%)	2 965 45.69%)
Number of probe (%)** associated with an island/shelve/shore			
hypermethylated	1 201 (34.79%)	1 842 (79.23%)	209 (26.06%)
hypomethylated	1 111 (32.18%)	114 (4.9%)	310 (38.65%)
outside of an island (Open sea)			
hypermethylated	176 (5.10%)	299 (12.86%)	190 (23.69%)
hypomethylated	964 (27.93%)	70 (3.01%)	93 (11.60%)

* (%) of all probes with p≤ 0.01 ** (%) of all probes with p≤ 0.001

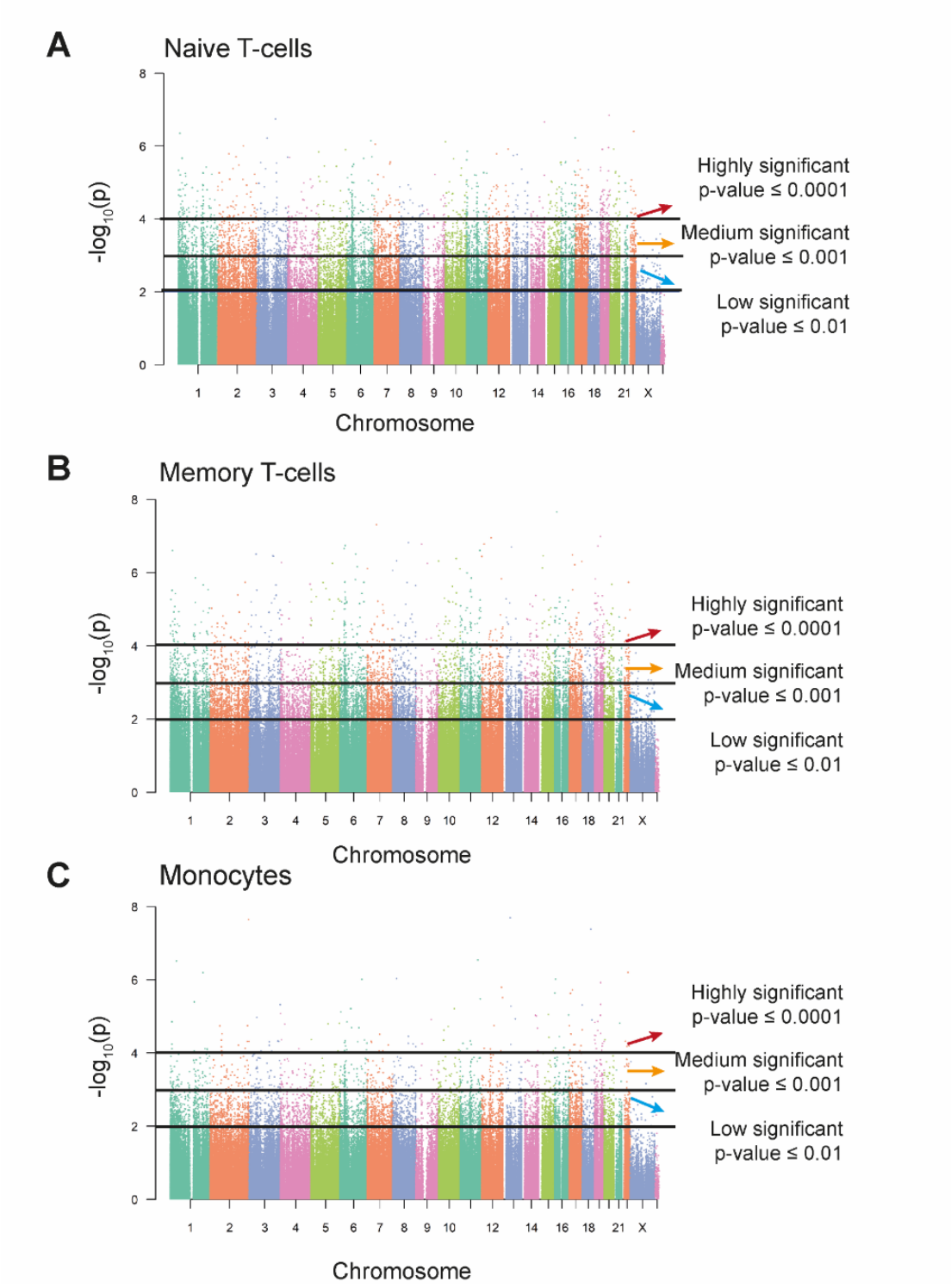


Figure 4-4 Manhattan plot displaying the -Log10(p-values) against the position on chromosome of ~480,000 individual CpG for A) naïve T-cells, B)memory T-cells, and C) monocytes. 3 thresholds of significance were derived from the plot (high, medium, and low).

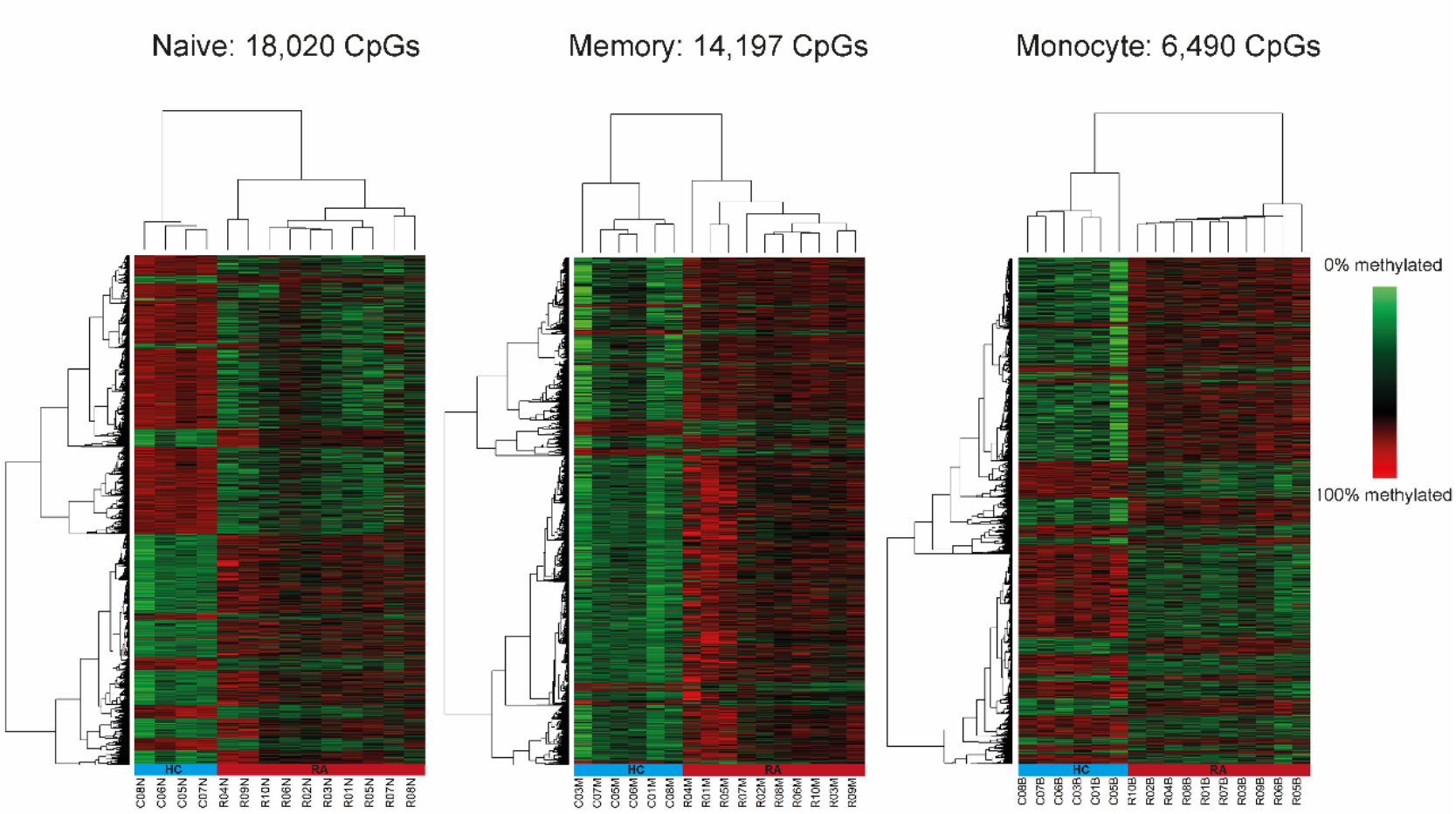


Figure 4-5 Hierarchical clustering and Heatmap representation of DM-CpG ($p<0.01$) between 10 RA patients (red bar) and 6 HC (blue bar) for A) naïve T-cells, B) memory T-cells, and C) monocyte. Note the high proportion of hypermethylation in memory T-cells (93%) compared to 46% in naïve T-cells and 54% in monocytes

4.3.1.3 Differential methylation patterns

To evaluate the possible effect of DM on gene structure, I manually inspected a few regions on the genome for the top p-value CpGs. Three typical patterns were identified (Figure 4-6);

- DM at a single CpG usually located in the core of a CpG island (example A),
- clusters of DM-CpGs in the shore/shelf of an island (example B),
- an isolated CpG in open sea (not always associated with a specific gene) (example C).

Example A illustrates the effect of the most significant DM-CpG observed (CpG06292898; p=1.06e-17) which occurs on a single CpG located in a 321 bp-long CpG island with no other difference found for the other 7 probes located in the island. Although this probe showed a very clear difference in methylation between HC and RA and there are several binding sites for transcription factors (such as SP1, TFAP2c, PAX-5 or E2F1) localised nearby this CpG (although not directly overlapping), that could nonetheless be influenced by this methylation. It may or not have a substantial biological effect, as a single CpG may not produce sufficient physical effect on the structure of the chromatin to affect gene expression.

Example B illustrate the effect of 16 medium significant DM-CpGs (p-values ($0.0003<p<0.001$)) located on the edges of a 47bp CpGs island. They were clearly clustered between 2 regions covering about 900 bp on the left shore/shelve of the island and 150 pb on the right. The close proximity of many differential hypo-methylations suggests they could locally have a cumulative effect on the chromatin and affect gene expression.

Example C is an isolated, highly significant DM-CpG (p=5.78e-13) far from any recorded island/gene or the next CpG interrogated by a probe (over 10,000 bp away). An effect of this is difficult to understand and hardly supportive of a direct gene expression change, although other distal effects can be imagined such as contribution to an enhancer (which are not specifically targeted by this 450K array).

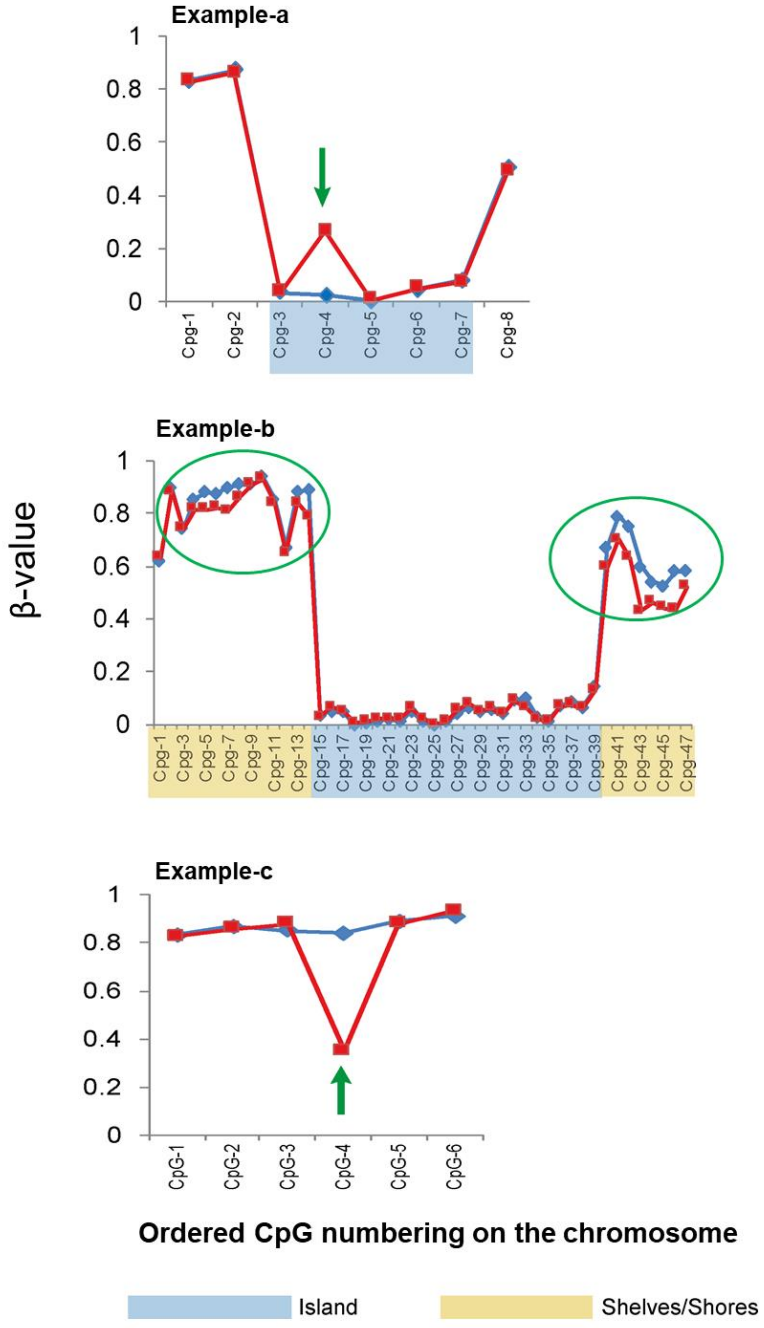


Figure 4-6 Examples of 3 typical patterns of the differential methylation. Example-a : DM at a single CpG. usually located in the core of a CpG island. Example-b : clusters of DM-CpGs in the shore/shelf of an island. Example c : an isolated CpG in open sea (not always associated with a specific gene).

4.3.2 Design of rules to prioritise clusters of differential CpG methylation

Considering the physical impact that DNA methylation could have on the chromatin packaging and DNA accessibility to the transcriptional machine, to continue my project, rather than relying only on the significance of DM between HC and RA for individual CpG loci, I decided to find clusters of significant differentially methylated CpG (DM-CpG-clusters) also called differentially methylated region (DMR).

I designed rules to define and prioritise DM-CpG-clusters and wrote a R code to analyse datasets automatically and systematically (R-code available on request).

The rules were designed with respect to the significance of individual CpGs and the distance (in bp) between them (Figure 4-7, A). Hypomethylation and Hypermethylation CpG/probe were investigated separately.

- The first rule was aiming at filtering only probes with a high significance for DM
  - **Rule1** : score = 1 if the p-value of the considered probe is highly/medium significant ($p \leq 0.001$)
- The second rule determined how many high/medium significant probes were located in a proximal region set within three nucleosome length (~750bp)
  - **Rule2** : score = number of probes with a p-value ≤ 0.001 and located within -/+ 750 bp of the considered probe.
- The third rule determined how many low significant probes were located in the same proximal region,
  - **Rule3**: score = number of probes with 0.001<p-value ≤ 0.01 and located within -/+ 750 bp of the considered probe.
- The fourth rule determined how many probes at all level of significance were located in a distal region, covering three further nucleosome lengths
  - **Rule4**: score = number of probes with p-value ≤ 0.01 and located between -/+ 750 bp to -/+1500 bp of the considered probe.

Further prioritising criteria were set. The probes which are highly/medium significant (p-value ≤ 0.001) and have at least one proximal high/medium significant probe (p-value ≤ 0.001) [Rule 1 score=1 and Rule 2 score ≥1] were selected. Lists of selected probes for all three cell types were generated. This resulted in low number of candidate probe obtained and no overlap for any probe was found between the 3 cell types. Therefore, that selection criteria may have been too strict.

More flexible filtering criteria were then set. This filtered probes that were highly/medium significant (p-value ≤ 0.001) and had at least one proximal significant probe at any level of significant (p-value ≤ 0.01) [Rule 1 score=1, and Rule 2 or 3 score ≥1]. The number of DM-CpGs-clusters with a score of >2 are listed in Table 4-3.

To further prioritize DM-CpG-clusters for a list of candidates for each cell subset, CpG were then prioritized using the sum of each rule : total score = score rule1+2+3+4). An example shows in Figure 4-7,B.

In monocytes, most clusters only showed the initial selecting CpG associated with only 1 other DM-CpG (final score=2). In contrast, in naïve T-cells, some clusters showed up to 7 selecting CpGs and scores up to 13, suggesting much larger effects on a wider region of the DNA. Intermediate results were observed in memory T-cells (score up to 9).

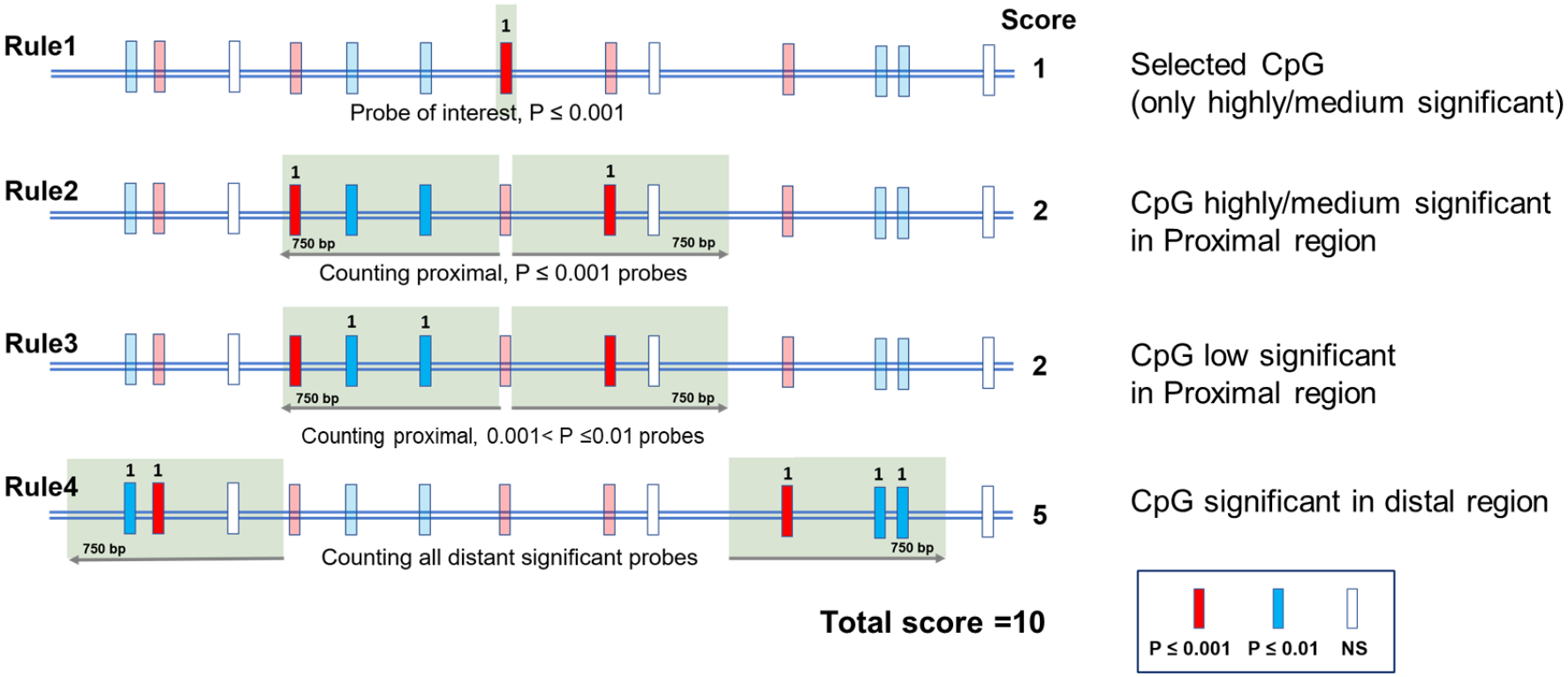
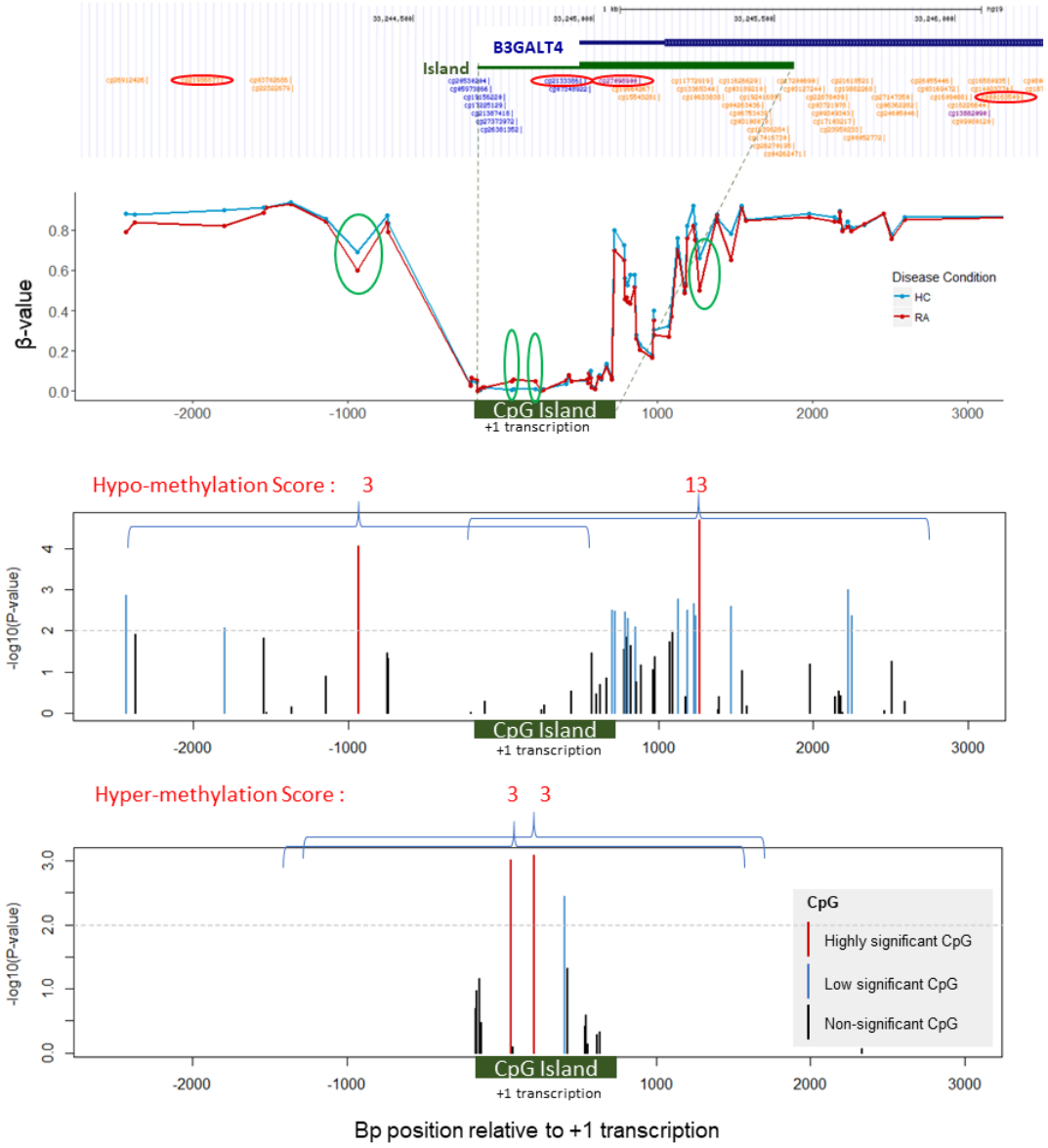


Figure 4-7 Design of rules to prioritise clusters of differential CpG methylation

A) Rule design for prioritising CpG. Rule-1 score 1 point vs. none, selecting only CpG-probes that had a high/medium significant p-value (represented by a red bar, p≤0.001). Probes of high/medium significance in a proximal region of +/-750bp (about 3 nucleosomes) were counted in rule-2 (1 point for each probe). Probes of low significance (blue bars, 0.001<p≤0.01) were separately counted in the same proximal region as rule-3 (1 point for each probe). Finally, rule-4 counted probes at all levels of significance (red and blue) in a distal region covering a further +/-750pb (1 point for each probe).



B) Example of a CpG island scoring. The top panel describe the structure of the B3GALT4 gene (blue box) and its CpG island (green box). The position of individual CpG (as cg.x codes, orange and purple) are displayed below the island. The methylation profile in RA (red line) and HC (blue line) are displayed in the second plot. 4 highly significant DM-CpG are circle in green (2 hypomethylated (shores) and 2 hypermethylated (core). Score resulting from the selection of these 4 CpG (rule-1) are displayed in the bottom 2 panel where –log10(p-values) are displayed against the chromosome sequence centred on the +1 transcription of the gene. 3 of selected CpG scored 3 points due to the presence of 2 more significantly DM-CpG in the area while the 4th CpG score 13 points due to a dense cluster of DM-CpG surrounding it.

4.3.3 Clustered and isolated DM-CpG in the 3 cell subsets

DM-CpG-clusters were then identify using the scoring system. The lists of gene associated with these clusters were derived. This process generated 648 genes for naïve T-cells (detailed in Table 4-3) 354 for hypomethylation and 294 for hypermethylation, 605 genes for memory T-cells and 58 genes for monocytes.

Emphasising this is still an exploration of data (including the heatmap and Manhattan plots previously described), naïve cells showed the most DM compared to other cells types and a bias towards more hypomethylations, which seems reasonable in terms of biological meaning as naïve cells are prompt toward response resulting in their differentiation into a particular helper T-subset. Thus they are particularly receptive to epigenetically clues and many sets of gene are available to be turned ON or OFF by changing DNA methylation levels.

The gene that received the highest score from the scoring system was beta-1,3-galactosyl transferase-4 (*B3GALT4*) in naïve T-cells (score=13, hypomethylated DM-CpG-cluster). The 2nd and 3rd highest score for hypomethylated DM-CpG-cluster were *TNF-α*, which presented no DM in memory cells or monocytes, and the Src-homology adaptor (*ABI3*) gene, (also specific to naïve cells). For memory T-cells, most of the DM-CpG-clusters were hypermethylated (600 genes) vs hypomethylated (5 genes). This hypermethylation suggested gene silencing which may aligned with the reported anergic characteristic of memory CD4+T-cells in RA patients (299) although this would need to be confirmed. The DM-CpG-clusters show the lowest scores in monocytes, which may suggest that epigenetic modification is not a mean of regulating cell function in such short live cell type.

Although DM-CpG clusters are likely to affect gene activity, highly significant CpG which no other significant CpG nearby (isolated DM-CpG), were still included in further analysis as they may have an effect on gene activity if the CpG is located in a distant transcriptional factor binding site for an enhancer. Isolated-DM-CpGs ($P<0.0001$, detailed in Table 4-3) kept were annotated with gene symbol (via array annotation). 266 hypomethylated genes were related to these isolated-DM-CpG in naïve cells (top 3 associated genes being *FOXF1, PARD6B, MAFF*) and 133 hypermethylated genes (top 3: *CHGA, SKI, FSTL1*).

Several of the genes on these lists appeared familiar to known biological processes implicated in RA. We further used ranking based on the p-value of the t-test to prioritise these isolated-DM-CpGs. Similar data for memory T-cells and monocytes are summarised in Table 4-3. Full lists of DM-CpG-clusters and isolated-DM-CpG in all cell subsets are available in (Data S1-3).

Table 4-3 Summary of the prioritisation of clusters of DM-CpG and associated genes.

	Naïve CD4 T-cells (n= 4HC,10RA)	Memory CD4 T-cells (n= 6HC,10RA)	Monocytes (n=6HC,10RA)
Hypermethylation			
Score range	0-9	0-9	0-4
DM-CpG-clusters			
Number of CpG with a score≥3	143	305	6
Number of CpG with a score=2	277	414	15
Score≥ 3 mean	3.91	3.85	3.5
Corresponding Number of gene (all clusters)	354	600	19
Isolated-DM-CpG			
Number of gene associated with a CpG in:			
island/shelve/shore	121	249	34
in Open sea	12	18	14
Hypomethylation			
Score range	0-13	0-2	0-4
DM-CpG-clusters			
Number of CpG with a score≥3	197	0	15
Number of CpG with a score=2	223	7	33
Score≥ 3 mean	4.31	na	3.26
Corresponding Number of gene (all clusters)	294	5	39
Isolated-DM-CpG			
Number of gene associated with a CpG in:			
island/shelve/shore	139	24	28
in Open sea	127	17	13

I then analysed the distribution of genes that were common to the different cell types (Venn diagrams displayed in Figure 4-8). For the DM-CpG-clusters, only 1 hypermethylated gene, 4-aminobutyrate aminotransferase (*ABAT*), was common to all 3 subsets and none for hypomethylation, suggesting differential effect of the disease on each cell subset while suggesting that the effect common to all 3 in this unique gene may be very strong. I considered that this finding may be artefactual although, this gene is associated with 9 DM-CpGs in naïve cells, 6 in memory and 4 in monocytes (most $p<0.001$).

Hypomethylated CpG associated genes common to both T-cell subset were no more numerous with only 1 gene but more for hypermethylation with 72 genes. For the isolated-DM-CpG, a similar analysis suggested again no overlap for either hypo or hyper-methylated gene between the 3 subsets.

Limited overlap was also observed between the T-cell subsets. The lack of commonality in methylation (only 1 gene) in 3 cells subsets and limited commonality between T-cell subsets points to the uniqueness of disease associated methylation change in each cells subsets in RA.

The most immediately recognisable DM genes associated with RA were cytokines/receptors including TNF/TNFRs, some IFN-signalling related genes, HLA-related genes, STAT family, and some integrin that has been known to involve in RA (Table 4-4). DM cytokine/receptor were numerous in naïve T-cells; several were repeated in memory T-cells and monocytes.

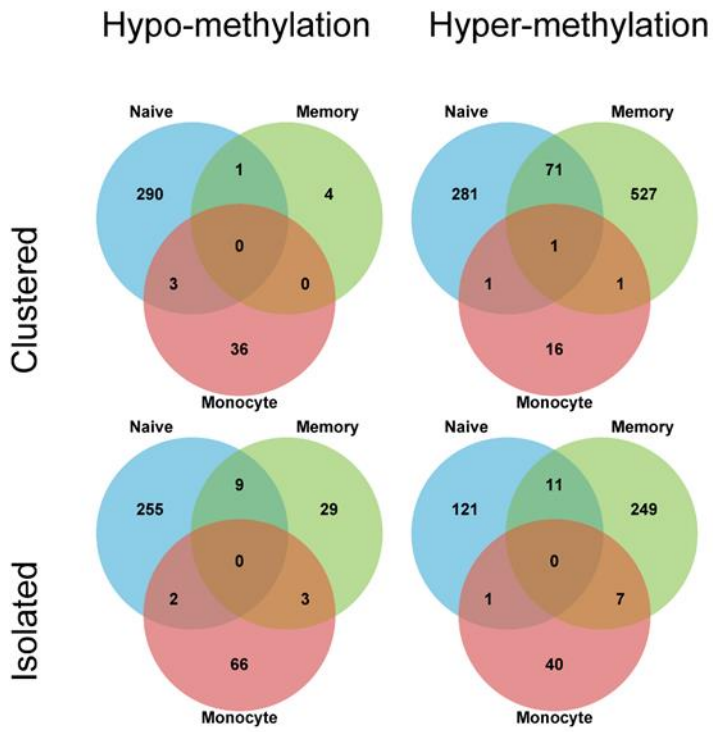


Figure 4-8 Venn diagram displaying the overlap between cell subsets for clustered and isolated DM-genes, hypo or hyper-methylated. Number in each section represent the number of overlapping genes**.**

Table 4-4 DM of cytokine genes in early RA

Gene Symbol	naive cells	memory cells	monocytes
INTERLEUKIN FAMILY			
high significance	*IL1B, IL31, IL2RA IL6R, IL21R*		
medium significance	*IL13,IL16, IL24, IL34, IL1R2, IL2RB, IL10RA, IL17RC, IL17REL, IL18BP, IL1RAPL1*	*IL6, IL17REL, IL17RA IL20RB*	*IL1RN IL12A*
low significance	*IL6, IL10, IL12A, IL15, IL17C, IL17D, IL19, IL21, IL25, IL36G, IL1RN, IL4R, IL15RA,IL17RA, IL20RB, IL21R, IL21RAS1, IL27RA*	*IL1B, IL12A, IL15, IL17D, IL24, IL37, IL1R2, IL4R, IL6R, IL12RB1, IL15RA, IL21RAS1, IL17RD*	*IL16, IL37, IL17RC*
TUMOUR NECROSIS GROWTH FACTOR FAMILY			
high significance	*TNF, TNFSF10, TNFRSF1B*	*TNFRSF10C*	*TNFAIP2*
medium significance	*TNFSF11, TNFSF14, TNFRSF8, TNFAIP3, TNFAIP8 TNFRSF13B,*	*TNFRSF19 TNFRSF13B, TNFRSF13C*	*TNFRSF1B*
low significance	*TNFSF18, TNFAIP2, TNFAIP8L1, TNFRSF1A, TNFRSF6B, TNFRSF9, TNFRSF10B, TNFRSF10C, TNFRSF18, TNFRSF19, TNFRSF21*	*TNFAIP8, TNFAIP8L3, TNFAIP8L1, TNFRSF9, TNFRSF10B, TNFRSF18, TNFRSF25*	*TNFAIP3, TNFRSF18, TNFRSF19, TNFRSF25*
INTERFERON FAMILY			
high significance	*IFNGR2*		
medium significance	*IFNL4*		
low significance	*IFNA2, IFNA7, IFNG, IFNGR1, IFNAR1, IFNAR2*	*IFNGR2, IFNGR1,*	
TRANSFORMING GROWTH FATOR FAMILY			
medium significance	*TGFA, TGFB2AS1, TGFB1I1*	*TGFB3 TGFBR3L*	*TGFA*
low significance	*TGFB1, TGFBI, TGFBR1, TGFBR2, TGFBR3, TGFBR3L*	*TGFB1, TGFBR2, TGFBR3,*	*TGFB1, TGFB2AS1*

4.3.4 Validation of DM gene

After establishing lists of DM-gene, the next step would be the validation of this data. I wanted to confirm that DM-genes identified on the Illumina methylation array dataset really have altered methylation levels, and/or whether such changes are able to affect downstream process such as transcription into mRNA or expression of protein.

4.3.4.1 Bisulfite sequencing of the TNF gene promoter in CD4+T-cells

Some of the genes on the DM-list above were already shown to be DM in RA in published research using bisulfite sequencing (228) and pyrosequencing (224), although not in early, drug naïve RA. These notably included TNF(300), IL6, IL6R (227, 228, 301).

I chose to use the bisulfite sequencing technique to validate DM of a CpG island. I selected the *TNF-α* gene given its high score (second highest score from the scoring system) and its central and unequivocal role in RA pathology, for further validation in early drug naïve RA patient (**Error! Reference source not found.**, A). This work was done in collaboration with a Master student in the group (co-author on the publication).

A 273 bp regions (indicated by a grey box on the *TNF* gene structure on Figure 4-9, B) in the *TNF* gene promotor encompassing 8 CpGs (3 of which present on the illumine 450K array) was amplified by PCR and sequenced from DNA isolated from total CD4+T-cells DNA (average purity of the CD4+T-cells population 97.5%) following cell sorting using magnetic beads from 7 HC and 9 RA patients. Details of the patients used in this study are included in Table 4-5.

Table 4-5 Demographic and clinical data for the control and RA patients used in the *TNF* bisulfite sequencing

Cohort 2 : bisulfite sequencing	**HC (n=7)**	**RA (n=9)**
age (years)*	55 (48-63)	46 (31-65)
M/F	1/6	3/6
ACPA (Pos/Neg)	na	7/2
Duration (months)*	na	15 (5-24)
TJC	na	5 (3-18)
SJC	na	3 (1-6)
CRP	na	10 (<5-83)

Data are presented as the median (range)

Hela cell control DNA, used as reference, was fully methylated at each of the 8 CpGs. In HC (n=7), I observed on average about 50% methylated/de-methylated CpG at each position, suggesting 2 populations of cells, one with methylated and one with un-methylated DNA of about equal proportion (Figure 4-9, A). In early, drug naïve RA patients (n=9), the overall proportion of cells with un-methylated DNA reached on average 90% (i.e. 10% of cells with methylated CpGs) at all 8 CpGs suggesting that most CD4+T-cells have altered their TNF gene, early in the RA disease process.

Looking back at the Illumina methylation dataset in this region of the *TNF* gene promoter (Figure 4-9, B, region in amount of the +1 of transcription), naïve CD4+T-cells showed partial demethylation with an average β-values of 50% methylation). DM was observed in RA with consistent hypomethylation of the whole region with on average a 7.1% less in β-values (range -2.3% to -20.8%). In contrast, this region is almost fully demethylated (average 22% β-values) in memory cells, while in monocyte the region was also fully demethylated (average 8% β-values) with no significant difference between HC and RA for both cell subsets.

The DNA methylation level of ~45%in total CD4+T-cells DNA observed by bisulfite sequencing is the result of the contribution of the methylation level from both naïve and memory CD4-T-cells. While there was no DM between HC and RA in memory CD4-T-cells, our data, confirm that in RA a large proportion of naïve T-cells have hypomethylated the *TNF* gene-promoter compared to HC, resulting in 90% of demethylated DNA in that region in a total CD4+T-cell DNA sample.

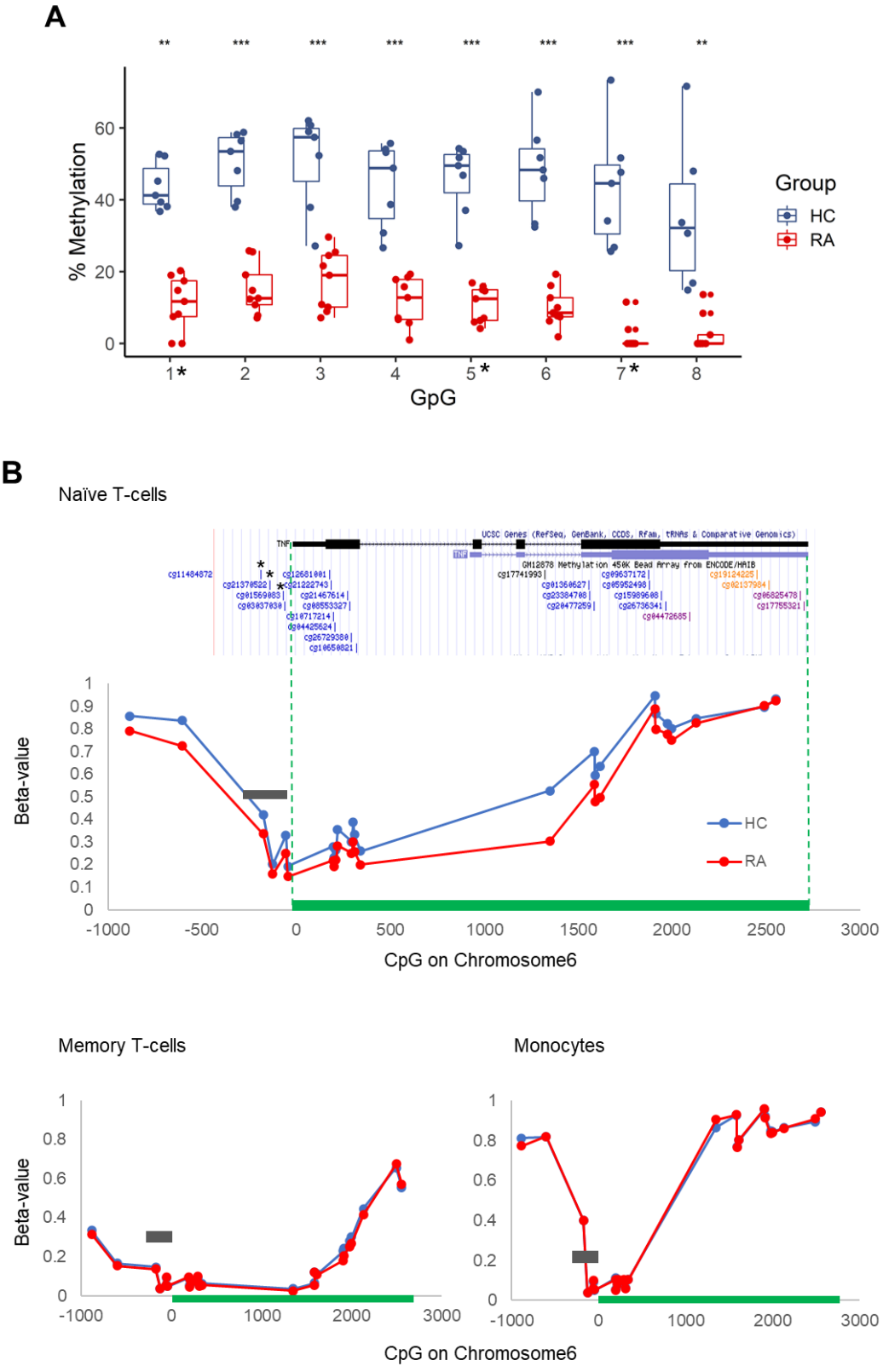

A
% Methylation
GpG
Group
HC
RA
B
Naïve T-cells
Memory T-cells
Monocytes
Beta-value
CpG on Chromosome6

Figure 4-9 DNA bisulfite sequencing of the *TNF-α* promoter region
A) A region of 273 bp in the promotor of *TNF* gene was selected for direct bisulfite sequencing of DNA extracted from total CD4+T-cell from HC and early drug naïve RA patients. Results of the sequencing covering 8 CpG (including the 3 used by the illumine 450K array highlighted by (*)), showed on average ~45% methylation in HC (n=7, blue dots) reduced to ~10% methylation in RA (n=9, red dots). Box plot represent the median and 25%/75% of the distribution of % of DNA methylation in each group. Statistical analysis comparison of HC and RA were performed using the MWU test (**p<0.01, (***p<0.001).

B) The TNF gene structure is presented at the top, showing the gene (black box) with bp numeration starting at the +1 of transcription (green box). CpG density in the Illumina 450K array are displayed below. 3 Illumina CpG included in bisulfite sequencing are highlighted by (*). Methylation levels (β-values) of CpG associated with the *TNF* gene obtained from the array results for naïve, memory CD4+T-cells and monocytes, plotted in order along chromosome 6. The region sequenced (273 bp- dark grey box) was located in the gene promoter. The median β-values of most CpG showed significant hypomethylation in RA patients (red line) compared to HC (blue line) in naïve T-cells. This region was highly de-methylated in memory cells and in monocytes although with no DM observed between HC and RA patients.

4.3.4.2 Differential gene expression compared to differential gene methylation in CD4+T-cells

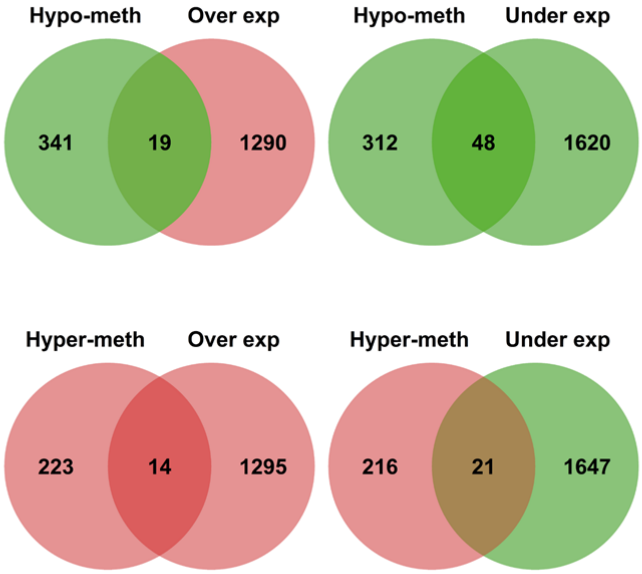
To validate the downstream effect of DM between HC and RA patient in mRNA level (302, 303), I compared the DM-gene list to a differentially expressed gene list obtained from similar early, drug naïve RA patients. This work was done in collaboration with another student in the group (co-author on the publication). We selected two publicly available gene expression datasets for CD4+T-cell (as there is no data available for naïve CD4+T-cells) from early, DMARDs naïve RA patient and HC (302, 303). After normalisation and aggregation of the 2 datasets (detailed is in the Appendix 6), we obtained a list of differentially expressed genes (DEG) (with adjusted p-value ≤ 0.05, FDR ≤ 0.05, fold change ≥1.5) between HC and RA. These genes included *JAK1,* TNF-family*, ICOS, CD69,* several MAP-kinases and their regulators, *TGF-β1, c-FOS* and *JUN*, HLA-related molecules, several IFN-signalling genes (*IRFs, IFITMs*), some TLRs, cytokines/chemokines their receptors and *PADI4*. These were aligned with the original findings published with the datasets (302, 303).

From the lists of DM-genes (DM-CpG-cluster with score ≥3 and highly significant isolated-DM-CpG (p≤0.0001)), 102 DEG genes could be matched using gene symbols after removing microRNA, open reading frames and other ambiguous gene symbols. These included 33 over-expressed and 69 under-expressed genes in early RA.

Amongst the hyper-methylated genes, *JAK1*, transcription factors (*TOX2, ZNF683*), cytokine (*IL12A*), growth factor receptor (*FGR1*), adhesion protein (*ITGA4*), ubiquitination (*FBXL3*), galactose transferase (b*4GALT2*), chromatin structural proteins (amphoterin/*HMGB1* and Protein *AF10/MLLT10*) were the 10 top over-expressed genes, while heat shock protein (*HSP72*), transcription factor (*ZNF213, ZNF219*), regulator of transcription (*TORC1/CRTC1, LMO4*), AP1-signalling (*AP1S2*), signalling adaptor (*SHKBP1, WLS, ASAP1*), ubiquitination (*UBAp2L*) and galactose transfer (*b3GALT4*), kinases (*STK10, PRKCD*) were under-expressed as summarised in Figure 4-10.

Hypo-methylated genes were matched for over expression with *IL2RA*, interferon signalling (*IFITM1*), phosphatases (*PTPRC, PSTPIP1*), as well as for under expression with *STAT5A*, kinases (*MAP3K11, CSK, TRIB1*) adaptor of signalling (*SH3BP4, CISH, SH2B2*), interferon response gene (*IRF8*), transcription factors (FGD2, TFEB), transporters (*SLC1A5, SLC16A3, SLC43A2*), regulator of apoptosis (*DAXX, CORO1A*). Taking the top genes based on fold differences in gene expression between RA and HC, the DM/DEG-genes associated with

known RA pathological pathways pointed to JAK1/STATs signalling, TNF-family, IFN-related signalling genes.



Hypo_Over 19 genes	Hypo_Under 48 genes			Hyper_over 14 genes	Hyper_Under 21 genes
ACSF3	*ABI3*	*HLA-DOA*	*TMEM94*	*B4GALT2*	*AP1S2*
BLOC1S2	*ACAP3*	*IRF8*	*TNFSF10*	*BRIX1*	*ASAP1*
C12orf10	*AP2A1*	*KCNN4*	*TRIB1*	*FBXL3*	*ASPSCR1*
CSTF2T	*ARHGDIA*	*LPAR1*	*TRMU*	*FGFR1OP2*	*B3GALT4*
FAM210A	*ARHGEF2*	*MAP3K11*	*TUBB1*	*GLRX3*	*BCL11A*
FBXW4	*B3GALT4*	*PGD*	*ZC3H12A*	*HMGB1*	*CRTC1*
GPRIN3	*BCKDK*	*PIK3CD*		*ITGA4*	*FBRSL1*
IFITM1	*CISH*	*PTPN6*		*ITM2B*	*FUOM*
IL2RA	*CLN8*	*SH2B2*		*MLLT10*	*FXR2*
LRRC32	*CORO1A*	*SH3BP4*		*RCAN3*	*HSPA1A*
PCED1B	*CSK*	*SLC16A3*		*SLC16A7*	*LMO4*
PCGF5	*CSRP1*	*SLC1A5*		*SMNDC1*	*PANX2*
PSMC5	*CTSH*	*SLC43A2*		*TMEM243*	*PEX11B*
PSTPIP1	*DAXX*	*SOX4*		*TOX2*	*PRKCD*
PTPRC	*EIF4G1*	*SPG11*			*SFT2D1*
SLC35F2	*ENGASE*	*STAT5A*			*SHKBP1*
SNORD32A	*FAM20C*	*SUSD1*			*SNX8*
SUCLG2	*FGD2*	*TFEB*			*STK10*
UBALD2	*FURIN*	*TLR9*			*UBAP2L*
	GPAT3	*TMC6*			*WLS*
	GRK6	*TMEM120A*			*ZCCHC24*

Figure 4-10 Gene expression analysis. Venn diagram and Table of overlapping genes between DEG (fold change expression ≥1.5, with adjusted p-value ≤ 0.05) and DM-genes (DM-CpG-cluster with score ≥3 and highly significant isolated-DM-CpG (p≤0.0001).

4.3.4.3 Cytokine expression compared to DM-genes

Many cytokines were present on the list of DM-genes (Table 4-4) including notably *IL1-β, IL17C, IL21, IL34 TRAIL, RANKL, LIGHT, TNF, TGF-β1* as well as many of their receptors (*IL2RA/RB, IL6R, IL10RA, IL17RC, IFNGR2, TACI*). Many of these cytokines were shown to be up-regulated at the protein levels in early RA patients (304-306); in particular IFN-γ, TNF-α, IL1-β, IL10, IL12, IL17, and IL6 for which there was also reports of association with DM at the gene level in established RA and at mRNA levels in PBMC cells from long lasting RA patients (228). There were DM genes on the list for 3 cytokines for which no data were available in early RA, IL21, IL34 and RANKL (for which the group since published data (307)). I selected these 3 cytokines for measuring protein levels in serum sample using 3 commercial ELISA. The serum samples were from independent HC (n=10) and early RA patients (n=20) who had closed demographic and clinical characteristics compared to the samples used for the illumine array (Table 4-6). This was due to the unavailability of matched serum sample and the benefit of having a larger samples size .

Table 4-6 Demographic and clinical data for the control and RA patients used in the Elisa.

Cohort 3 : ELISA	**HC (n=10)**	**RA (n=20)**
age (years)*	51 (40-63)	60 (41-75)
M/F	4/6	7/17
ACPA (Pos/Neg)	na	14/6
Duration (months)*	na	4 (1-24)
TJC	na	12 (5-22)
SJC	na	11 (0-22)
CRP	na	20 (<5-60)

Data are presented as the median (range)

All 3 cytokines showed higher levels in early RA (all $p<0.001$, Figure 4-11, A). Combined with data generated by my supervisor in previous published work (308) showing higher levels of IL1-β, IL6, IL10, IL12, IL17A, IFN-gamma and TNF-α (all $p<0.001$) in HC (up to 20 samples) and early, DMARDs naïve RA patients (up to 40 samples) (Figure 4-11,B), these data confirmed the possible effect of DM genes at protein level for these cytokines.

Altogether, both validations (DEG and ELISA) showed that many of DM-gene on the list, demonstrate associated change of expression at mRNA or protein levels, suggesting a pathophysiological contribution of such changes in DNA methylation in early RA.

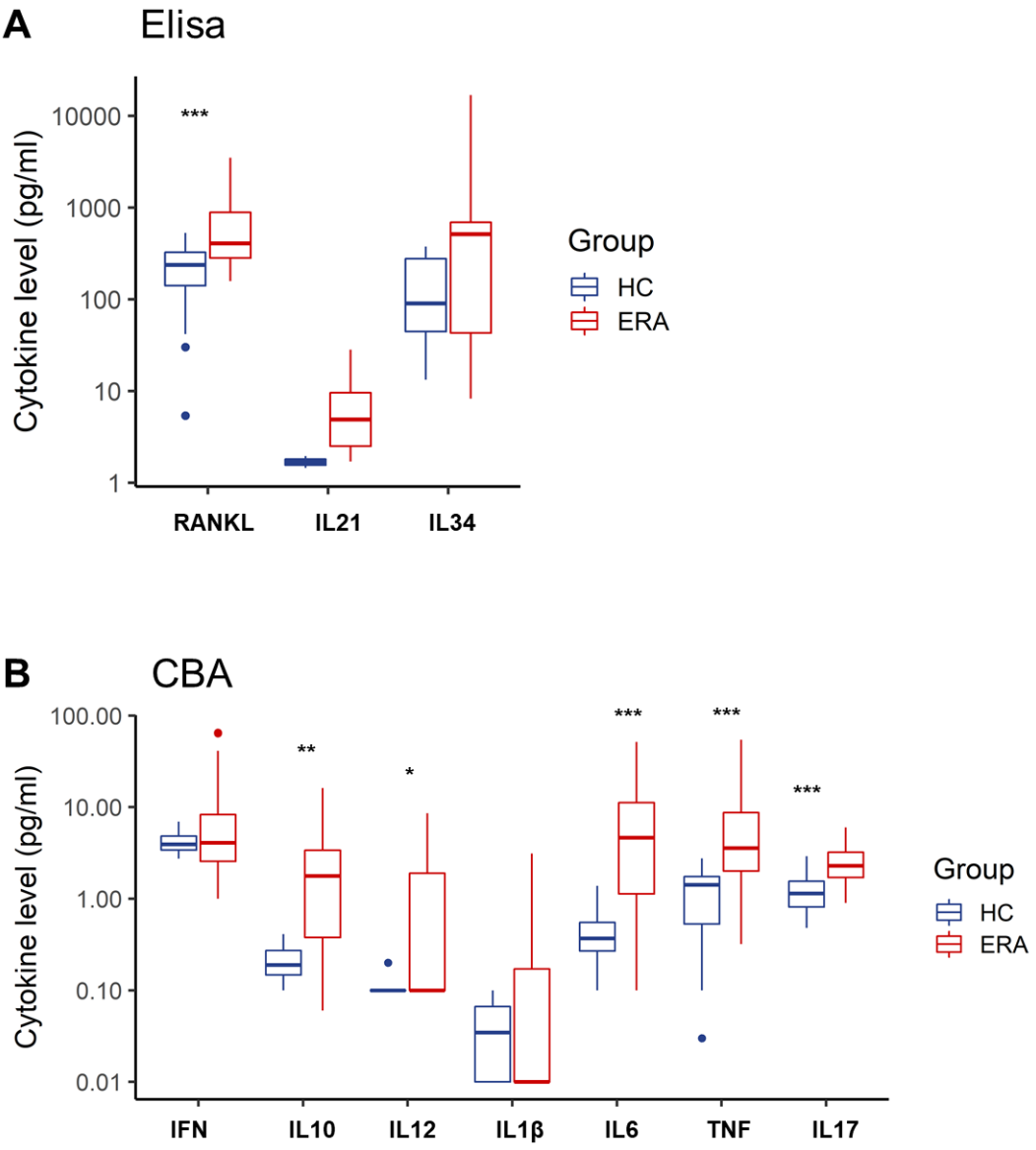


Figure 4-11 Levels of expression of several cytokines in HC and early RA patients.

A) Novel data are displayed. Serum samples were tested for cytokines levels using ELISA. The samples were collected from HC ((grey box plot in each pair, n=10) and early RA (black box plot, n=20). All cytokines were significantly over expressed in RA (Mann-Whitney U-test , p≤0.001).
B) Data recapitulated from previous studies (308). Data were generated by cytometry bead array or ELISA (unpublished) from variable number of samples from HC (up to 20) early drug naïve RA patients (up to 40).

4.3.5 In silico functional interactions between products of DM-genes in naïve CD4+T-cells.

I next explored whether the DM genes in naïve CD4+T-cells (DM-CpG-clusters (≥3) and isolated-DM-CpG/genes (p≤0.0001) including 591 gene symbols would point to specific pathways and/or functions that could be associated with pathogenesis. I constructed hypothetical functional networks using the STRING database (279, 280) of known and predicted protein-protein interactions which mines direct and indirect physical interactions and/or functional associations from several knowledge-databases, including genomic context, laboratory experiments, co-expression and text mining. The 70% confidence was used in interactions setting.

I generated an initial network with several association/interaction nodes (Figure 4-12, A). The network was shaped around interaction of several nodes centred on cytokines (TNF-α, IFN-γ, IL1-β and TGF-β). This software suggested a manual addition of several genes to strengthen nodes in the proposed STRING interactions. This notably included genes from the cytokine/receptor list (Table 4-4, medium and lower significance) as well as intermediate kinases and downstream signalling molecules in cascades. I interrogated methylation data manually for all suggested genes. I included or rejected suggested genes based on whether they had a DM-CpG cluster with a score=2 or an isolated-DM-CpG (P≤0.001).

The genes suggested included several members of the IL17/IL17R axis. Th17 cells develop from naïve T-cells resulting in full polarisation in memory T-cells. IL17 related genes were therefore added to strengthen this axis also because they were showing clear DM in memory T-cells. Other additions were accepted for isoforms of regulators of cytokine signalling (SOCS), caspases and regulators of apoptosis and finally, additional members of the interferon related gene signalling cascade. The final gene list (n=687) is detailed in supplementary files (Data S4), detailing all manual additions (96 genes).

At the end of this analysis, my final STRING model (Figure 4-12, B) clearly displays an interaction network centred on several JAK1/STATs nodes, and defining several groups of genes :

- a 1st linked to IL6/IL6R/STAT3 signalling (blue group of genes),
- a 2nd for IL27RA/STAT2 (orange group) linking to downstream interferon signalling,

- a 3rd centred on IL2/IL15/STAT5 (dark blue group)
- and a 4th on IL12/IL13/STAT4 (green group).

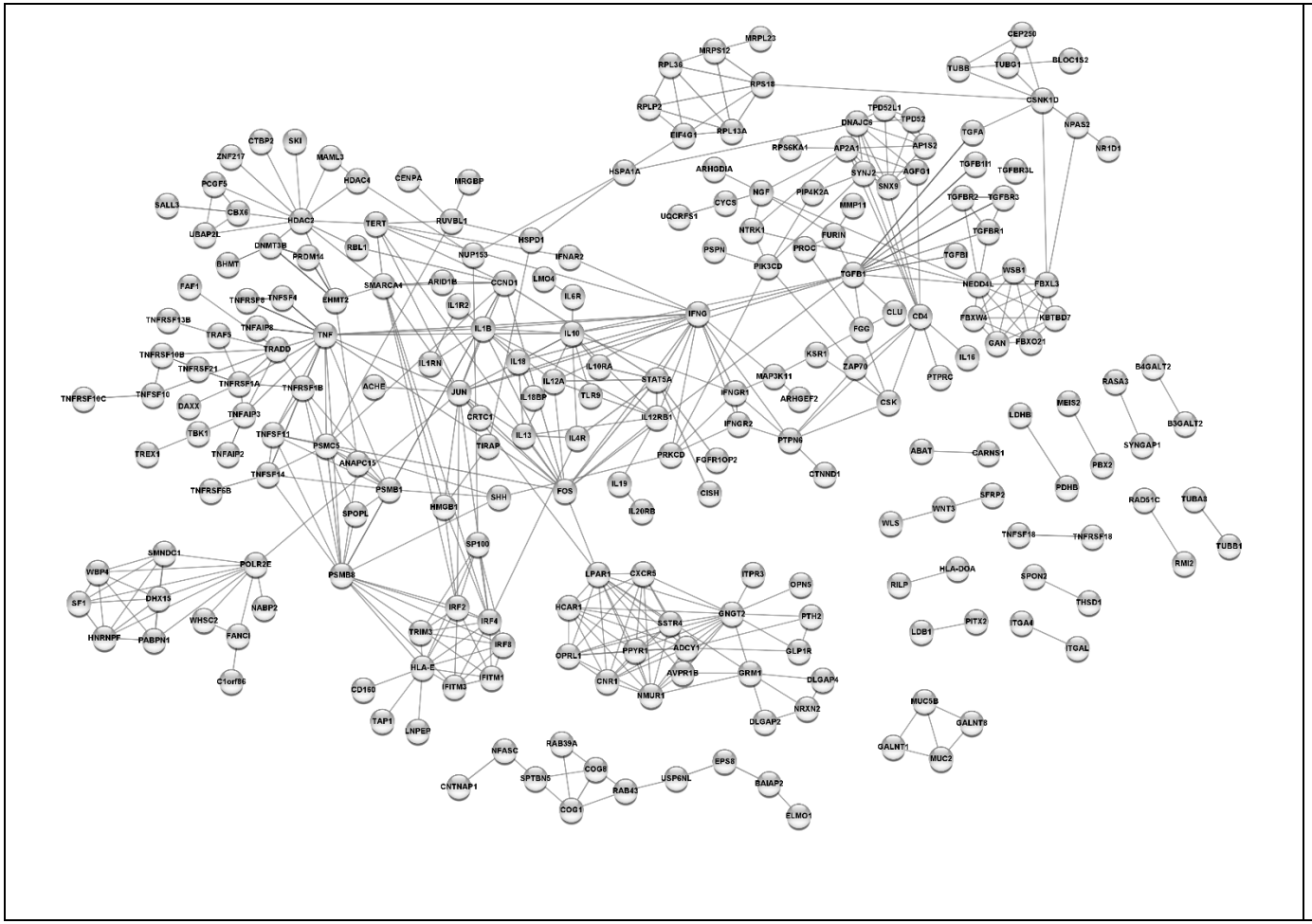
Many other associations were suggested, notably with respect to

- TGF-β signalling (pink group).
- an IL17/IL17R related group of genes (duck green group)
- as well as a one involving epigenetic programming (yellow group).

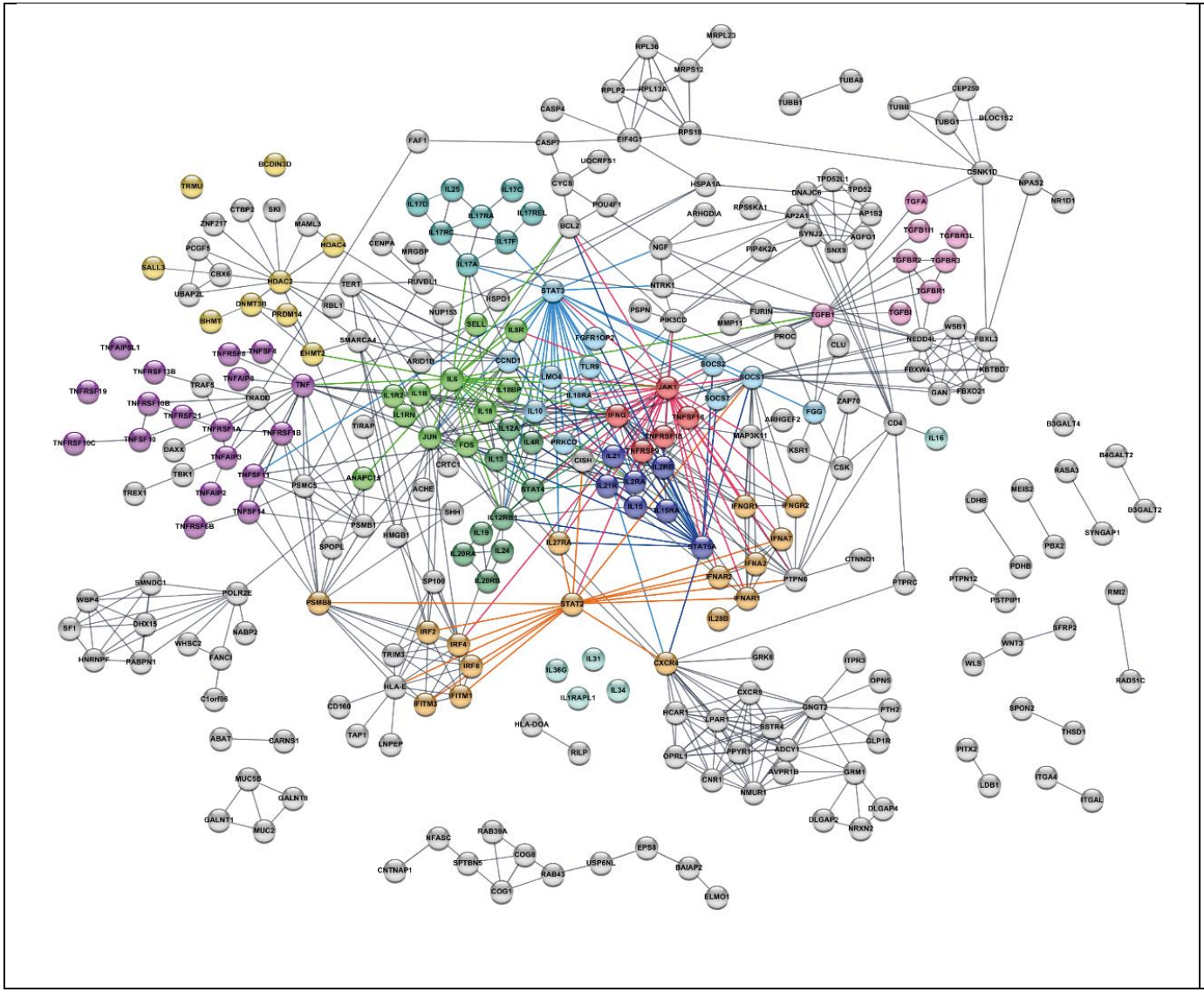
Many of the links (grey lines) were associating proteins with IL6, which itself was linked to a

- TNF-α signalling (purple group).

Figure 4-12 Functional interaction network (STRING analysis)



A)The initial STRING model included genes symbol/proteins from the DM-cluster-CpG with a score ≥3, and highly significant isolated-DM-CpG for both hypo- and hyper methylation.



B) The Final STRING model included the same genes symbol/proteins from DM-CpG, and manually added genes as suggested by the program when verified for cluster scores =2 or medium/high significant isolated-DM–CpG.

The network analysis displays JAK1/STATs nodes link to several signalling pathway :

- JAK1 signalling node in red;
- STAT3 node (blue) linked to
- IL6 signalling (light green);
- STAT2 and downstream interferon related signalling (orange);
- IL15/IL2/STAT5 node (dark blue);
- IL12/IL13/STAT4 (green);
- TGF-B signalling node (pink);
- IL17/IL17R axis (duck green);
- TNF-α related genes (purple);
- DNA methylation/modification related genes (yellow).

I subsequently analysed the IL6 and TNF signalling cascades in more details (Figure 4-13, done in collaboration with another student in the group, co-author on the publication). The component of the IL6 and TNF signalling cascades were retrieved form several data knowledge sources and assembled into a signalling cascade schematic. The DM-genes (DM-CpG-cluster score>=2, isolated-CpG p-value <0.001) were mapped to these schematic with green for hypo-methylation and red for hyper-methylation labels (Figure 4-13, A and B). Many DM-genes were directly involved in the TNF signalling cascade while similar and/or additional genes showed the same for the IL6 pathway.

I also aligned DEG (with adjusted p-value ≤ 0.05, fold change ≥1.5) using the previous analysis (as described on page 106) onto these two signalling cascades (Figure 4-14, A and B, red as over-expressed and green as under-expressed in RA). This highlighted *JAK1* and *STAT3* and *4* and many more genes, which, following manual investigating into DM, also showed low/medium significant levels of DM at the DNA level which was not prioritised.

Altogether, this work tied together many observations and focalised my investigations further into IL6 and TNF as central to early disease events.

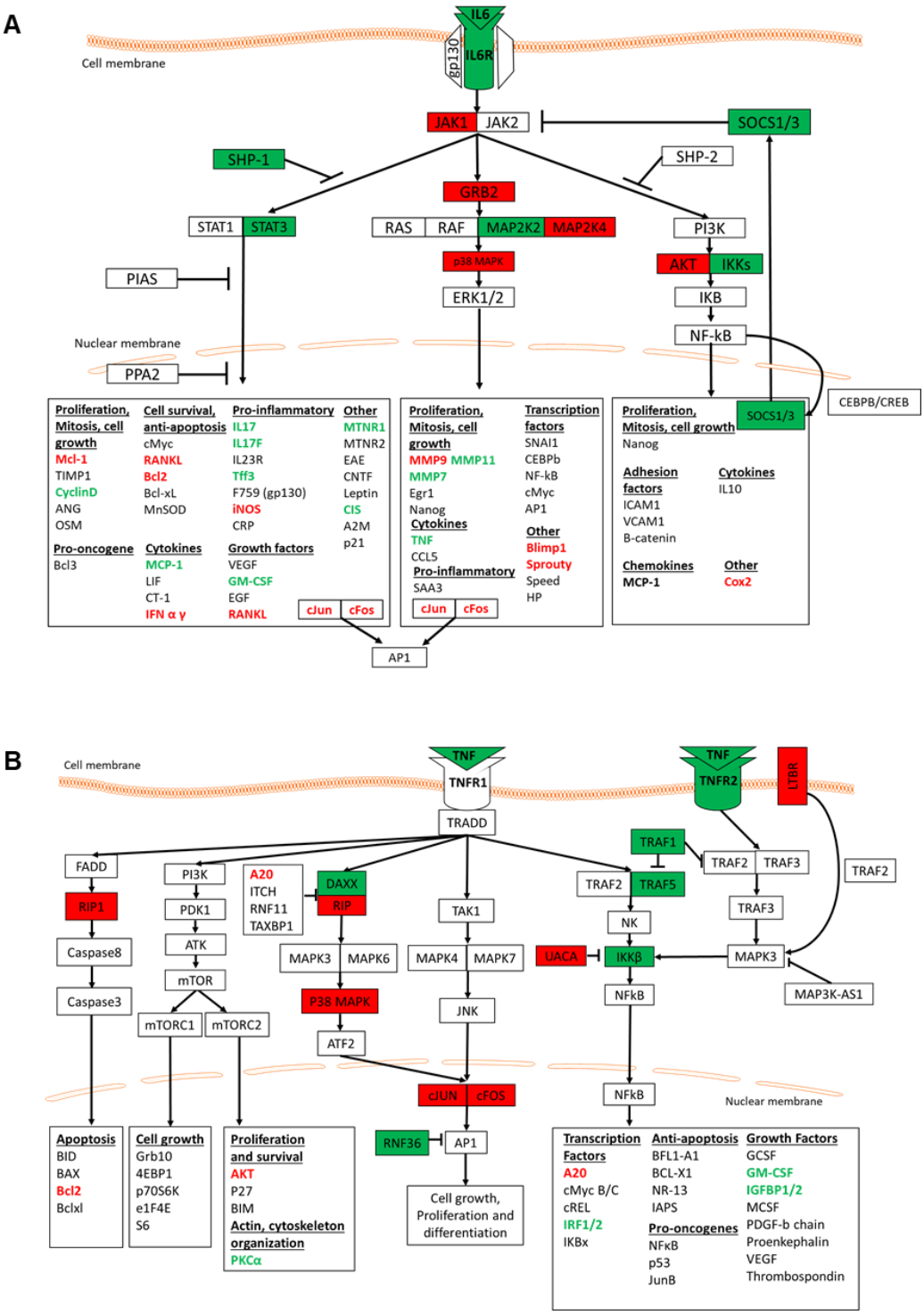


Figure 4-13 DM-genes in signalling cascade for a) IL6 and b) TNF-α. Components of the signalling cascades were listed from several data knowledge sources, and assembled as a signalling cascade schematic. Hypomethylated DM-gene in RA are represented in green; hypermethylated DM-gene in red.

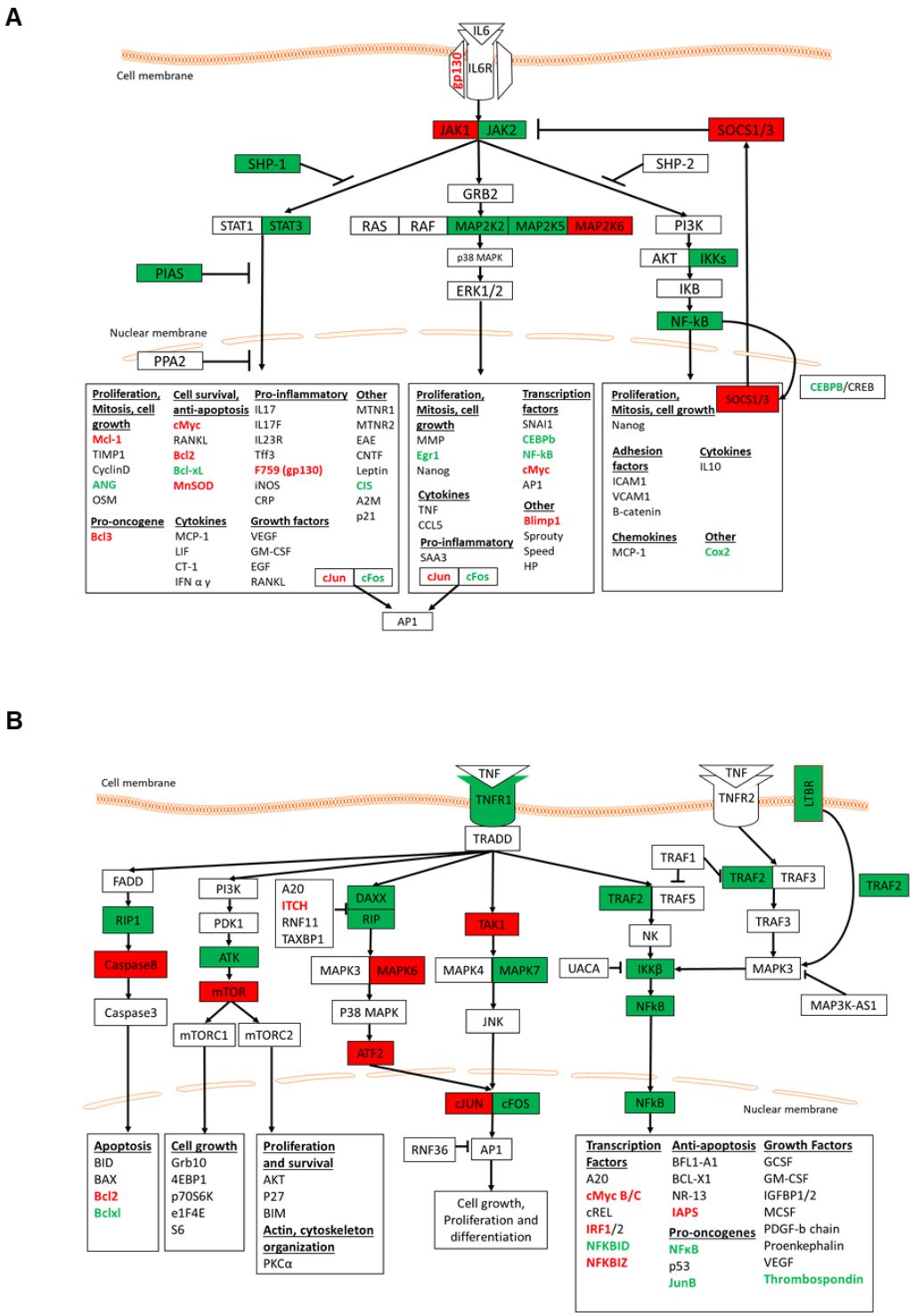


Figure 4-14 DEG-genes in signalling cascade for a) IL6 and b) TNF-α. Components of the signalling cascades were listed from several data knowledge sources, and assembled as a signalling cascade schematic. Under expressed-gene in RA are represented in green; over expressed-gene in red.

STRING network analysis of memory T-cells and monocytes DM-genes.

A similar STRING analysis was performed on the memory and monocyte datasets. Hypermethylation was mainly observed in memory T-cells, suggesting a potential global gene silencing effect (by analogy to the cancer field). The STRING analysis of DM genes (DM-CpG-clusters (score ≥3), isolated-DM-CpG/genes (p≤0.0001), and some manually add gene if score ≥2 or p≤0.001) in a list of 502 gene symbols, suggested a single major network, centred on *EP300* (Figure 4-15, A). This node then linked to several other genes, *SKG1*, histone modification enzymes (*Hist3H2A*), transcription factors (*HES5, PAX6, FoxO3*) and signalling protein (*TRAF6*).

EP300 encodes the histone acetyltransferase p300 that regulates transcription of genes playing an essential role in cell growth and/or differentiation, notably preventing tumor growth. EP300 contains a domain that recognizes acetylated lysine residues bromodomain that is known to be involved in IL6 signaling (309) and as a co-activator of hypoxia-inducible factor 1 alpha (HIF1A) resulting in VEGF induction. IL6 signalling therefore appears also central to the memory networks via EP300.

Serine/threonine-protein kinase (SGK1) regulates ion transport and is under the control of stimuli including insulin (as seen here with the insulin receptor gene INSR), growth factors and glucocorticoids (310). It has been shown to contribute to several pathways including inflammation, cell proliferation and apoptosis (311).

A similar analysis run for monocyte specific DM-genes (from a list of 187 gene symbols) revealed no particular dominant node (Figure 4-15, B). Although *IL6R/IL6* were not themselves DM in monocytes, *EP300* remained central to the network generated for this subset.

Overall from the STRING network construction, the same number of DM genes in naïve cells (687 DM genes entered) brings up more relationship and highlights several pathways than in memory cells (502 DM genes). The network contains genes that are both unknown and well-known for being involved in RA or inflammation/immune responses. The known genes shown in the network confirm linkage with established knowledge while they also support forming new theories about early pathological events in RA (This point will be further mentioned in the discussion).

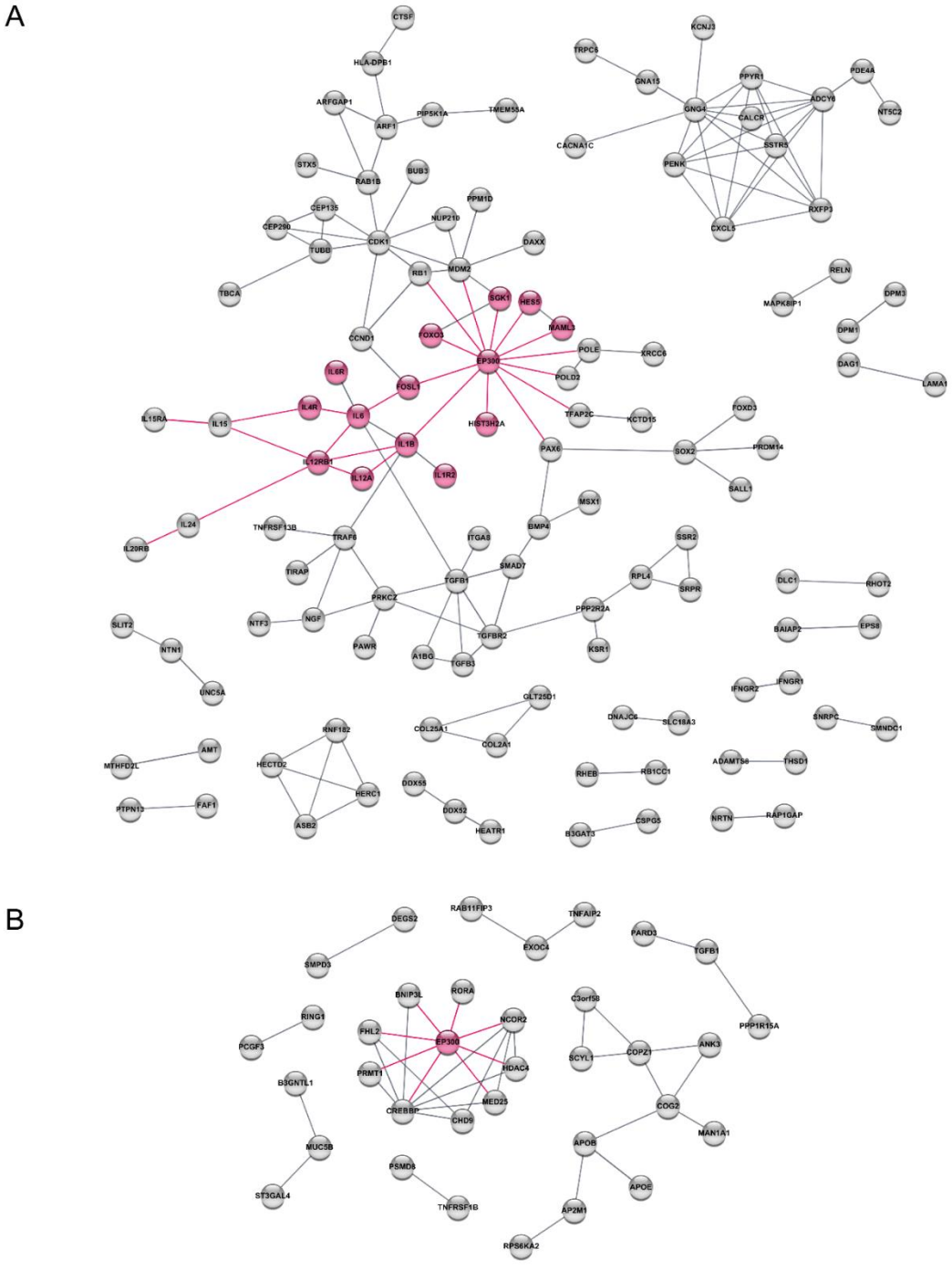


Figure 4-15 STRING network of functional relationships between DM-genes in A) memory T-cells and B) monocytes. The small network related to IL6 signalling and EP300 are heighted in pink.

4.3.6 Validation of the scoring system using available R-packages

My DNA methylation analysis workflow used standard procedures of data management, with an in-house analysis for prioritising DM-CpG-cluster and isolated-DM-CpG. DM-CpG-cluster were scored based on the physical positioning of CpG/probes, and a t-test for comparing methylation level between HC and RA. Therefore probe-poor regions may have been discarded due to lack of physical proximity in their positioning, rather than by lack of individual methylation different between HC and RA. I therefore also considered highly significant isolated DM-CpGs in the candidate lists. However, For the analysis of high-throughput genomic data, it is usually suggested to correct for multiple hypothesis testing, estimation of the false discovery rate (FDR) controlling the family-wise error. Controlling for multiple comparisons in our data left us with very little significant CpGs the further analysis (most likely due to n=10), so I decided not to considered this in my primary analysis, this being an exploratory study whereby some false-positive results were an acceptable risk.

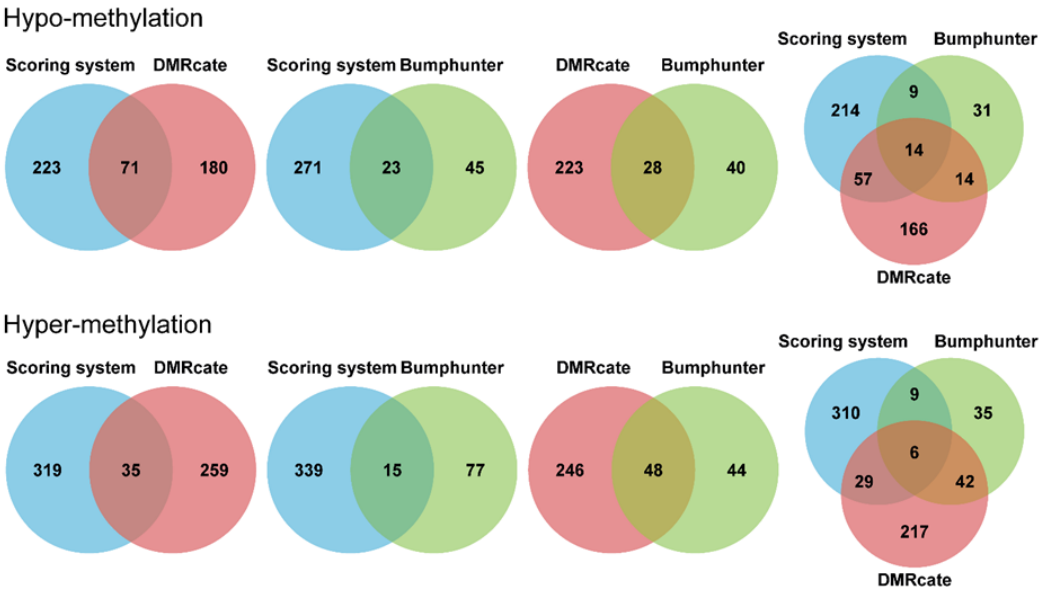
In order to further validate the scoring system, I ran the DNA methylation dataset through two publicly available R packages; DMRcate (273) and Bumphunter (272). These 2 packages aim to find a differentially methylated region (DMR), based on a similar concept to my analysis finding DM-CpG-Cluster, although they use more stringent statistics.

DMRcate uses moderated t-statistic to find significant individual CpG and then uses a function to agglomerate the nearby region from groups of significant CpG (273). It allows for several parameters to be adjusted such as the statistical significance (FDR adjusted p-value), the methylation difference itself (the β-value difference between group), and the length of the region to consider. The DMRs can then be annotated to the associated gene. Running this analysis using the recommended/default settings (FDR ≤ 0.05, 1000 bp, average Dβ-value ≥0.05), only 2 genes were DM in *RA, CUTA* (CutA divalent cation tolerance homolog), and B3GALT4 (Beta-1,3-Galactosyltransferase 4). Using customised settings with similar criteria to my scoring system (allowing results to be compared) resulted in a list of genes including 251 hypo and 294 hyper-methylated genes respectively (Data S5). When compared to my scoring system list in naïve T-cells, results showed a large overlap with 106 genes commonly identified by both analyses (Figure 4-16).

Bumphunter finds regions of interest based on regression modelling and smoothing techniques as well as permutation to assess the statistical uncertainty for each DMR (272). I attempted to run this package despite our small number and obtained a list of “bumps” with 68 hypo and 92 hyper-methylated genes respectively (Data S5) These were overlapping with my scoring system for 38 genes.

An overlap between the three strategies revealed 20 genes (14 hypo and 6 hyper-methylation) in common (Figure 4-16). The most immediately recognisable genes were *TNF*, some IFN-signalling related genes; *IFITM*, *PSMB9* and *AIM2*, and some transcriptional regulator; *MEOX1, EOMES* and *HIC1*.

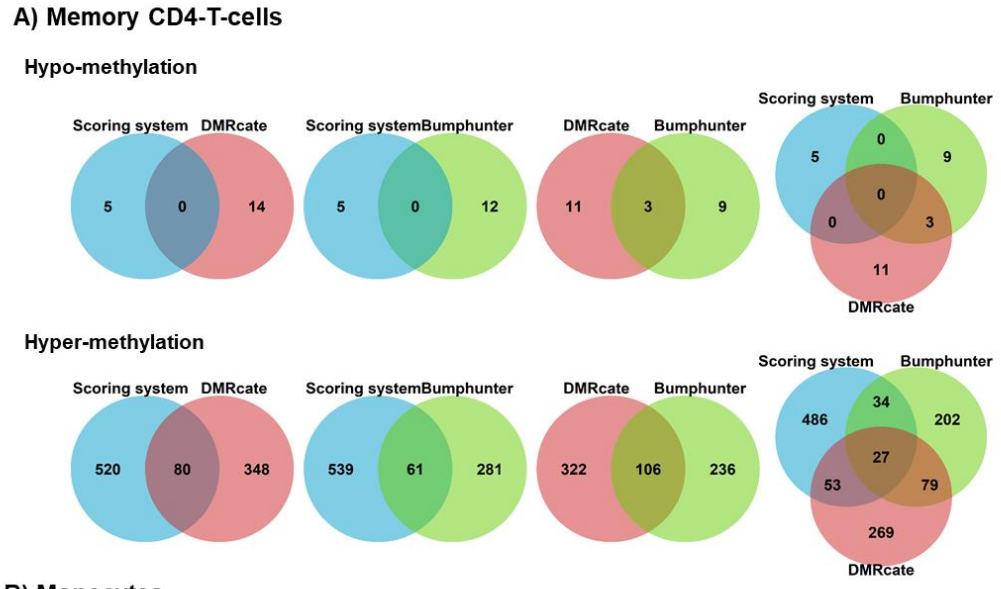
A similar analysis was performed for memory T-cell and monocytes subset (Figure 4-17, A and B). 27 hypermethylated were overlapped between 3 strategies in memory T-cell while 4 (3 hypo and 1 hyper-methylation) overlapping genes found in monocytes. Interestingly, *ABAT* remained though for all strategies in 3 cell types.



3 Packages overlaping DM genes

Hypo-methylation 14 genes			Hyper-methylation 6 genes
ABI3	*IRX4*	*SORBS2*	*ABAT*
AIM2	*KSR1*	*TNF*	*EOMES*
CORO1B	*LPIN1*		*HIC1*
HNRNPF	*MEOX1*		*MAD1L1*
IFITM1	*NEAT1*		*PRRT1*
IFT140	*PSMB9*		*ST8SIA6*

Figure 4-16 Overlapping of naïve T-cells DM gene between 3 strategies. The DNA methylation data of naïve T-cells were analysed using our scoring method, the DMRcate and the Bumphunter packages. Overlap between list of DM-genes are presented as Venn diagrams and gene symbols (Table).



Memory CD4+ T- cells

Hyper-methylation 27 genes				
ABAT	*DUSP6*	*M1AP*	*RSG1*	*ZIC5*
ADAM32	*GRIK2*	*MAD1L1*	*SDR42E1*	*ZNF577*
BHMT	*HIC1*	*MAJIN*	*SGK1*	*ZNF667*
CBX5	*HORMAD2*	*NFAM1*	*SOX1*	
CTNNA2	*KLF15*	*PHACTR2*	*TBPL2*	
DNAJC6	*LINC00461*	*POM121L2*	*TFAP2A*	

Monocytes

Hypo-methylation 3 genes	Hyper-methylation 1 gene
ECHDC3	*ABAT*
IER3	
PRMT1	

Figure 4-17 Overlapping of DM gene between 3 strategies. The DNA methylation data of A) memory T-cells and B) monocytes were analysed using our scoring method, the DMRcate and the Bumphunter packages. Overlap between list of DM-genes are presented as Venn diagrams and gene symbols (Tables).

4.4 Discussion

My study of genome-wide methylation analysis of T-cells helped gain more understanding of RA pathogenesis. Despite being an exploratory study it could both confirm the involvement of known pathway (mainly TNF, IL6 and type-1 IFN) in the early disease pathogenesis and importantly discover up new candidate genes or pathways.

The amount of methylation information offered also depend on the technology of choices. Here for this project, I use the 450K array, an improved version from 27k array, measuring ~485,000 CpG on the genome although aiming mainly at promoter regions and gene body, which covered most of CpG islands identified in the human genome, their shelf and shore but very few CpG Island outside these structures. The methylation information provide is therefore limited to the coverage and the probe design appropriate for this technology.

To validate results from my scoring concept I compared results generated with other analysis tools (DMRcate and Bumphunter). The overlapping number of DM-gene between these two strategies and my scoring system was not very large. This is likely due to the statistical design behind DMR finding algorithms and the strictness of the selected criteria. The main weakness of my scoring system was that it focused on statistical significance, not accounting for the actual difference in methylation level (delta β-value difference) which was included in the other two packages. Therefore, highly significance CpG could present with small methylation difference. Relaxing setting criteria naturally provides more DM genes candidates but can overload list including false positives. On the other hand, too strict criteria resulted in no candidate at all. A balance between these therefore needed to be met, and the aim of my exploratory analysis, using relaxed criteria, was therefore to generate lists of candidate genes that could lead to further understanding about early events in the disease pathogenesis. Indeed molecular and cellular functional investigation will be needed as further steps to definitely prove and explain how change in methylations of particular genes can explain the pathological mechanisms of the disease. Thus, my scoring system, which calculated a score on individual CpG (instead of picking up the whole region with high DM), will be useful in designing further study at a base-pair level to relate individual changes with functional effect. Another limitation of my scoring system (and also for the other packages) is that it focuses on regions with high CpG density. Genes with more probe will get better chances to be scored. Thus these are biased toward the chosen areas/genes coverage by the Illumina

probes. Thus the result might not be fully representative of the biological differences between HC and RA.

Despite the limitations of my scoring system, data provided lists of DM-genes which were partly validated by the other packages and highlighted several RA pathogenesis pathways. Combining this work with my supervisor previous work also pointed to the new insight into the RA pathogenesis which were published in a manuscript entitled “Differential CpG DNA methylation in peripheral naïve CD4+ T-cells in early rheumatoid arthritis patients” on 07 April 2020 in Clinical Epigenetics journal.

The initial exploration and analysis to identify DM gene in naïve CD4+T-cells, memory CD4+T-cells, and monocytes from of HC and early RA patient using my in-house scoring system showed that the DNA methylation alterations in RA for each cell types is quite unique. The very limited number of genes being commonly affected between in 3 cell subsets reinforced the fact that such epigenetic modifications are highly cell-specific, similarly to previously reported (312), and suggested different effect on each cell types in disease pathogenesis**.**

The only overlapping DM-gene to all 3 subsets was *ABAT*, 4-Aminobutyrate aminotransferase enzyme. This gene is responsible for the catabolism of gamma-aminobutyric acid (GABA), an essential inhibitory neurotransmitter, reducing neuronal excitability throughout the nervous system, and directly responsible for the regulation of muscle tone. T-cells express the GABA receptor and exposure to GABA is involved in regulatory loop reducing inflammation and promoting "regulatory" responses (313, 314). As such GABA has been implicated in autoimmune diseases in animal models including arthritis (315, 316). Hyper methylation in *ABAT* may therefore contribute to a loss of GABA’s regulatory effect in all blood cell types, promoting inflammation and autoimmunity. Being similarly regulated in all subsets, the DM of *ABAT* suggest a general effect on this particular gene, while cell subset specificity appears to be the rule for most others.

Methylation changes in RA were observed more in naïve CD4+T-cells than memory CD4+T-cells or monocytes. DM cytokine/receptor genes were numerous in naïve T-cells while several were repeated in the other two cells. Consider the involvement of T-cells in the early stage of RA pathogenesis and that naïve T-cell is an immature cell with an ability to differentiated to particular effector cells depending on the stimuli, it is sensible that the methylation alteration in RA were targeting naïve T-cells; especially on the cytokine/chemokine receptor genes which receive signals from the outside environment as well as for the cytokine

production itself which is a sign that naïve cells may be changing toward a particular polarisation subtypes.

Focusing on naïve CD4+T-cells as the most susceptible cell type to change in DNA methylation, a functional relationship network analysis of DM gene highlighted a central role for IL6/STAT3 with downstream effects on several pathways including TNF signalling, the potential for differentiation towards Th17 cells and interferon signalling.

IL6's importance is well recognised in RA (165, 301), as is the effectiveness of the therapeutic IL6 blockade (119). DNA methylation alteration of *IL6/IL6R* genes was reported in SF, PBMC, and T-cells of established RA patients (227, 228, 301). IL6 has been shown to induce the DNA methylation change in cancer and SLE (243, 317). Recent gene expression analysis in CD4+T-cells also suggested a role for IL6, very early in the RA disease process, notably compared to patients with inflammatory joint symptoms not progressing to RA (302). Increased IL6 levels in serum of early, drug naive RA is well established and was shown here using data from my supervisor (304-306). The functional effects of IL6 on CD4+T-cells have been explored extensively (reviewed in (318)). Specifically, in naïve CD4+T-cells, IL6 induces survival (319), proliferation (320) while memory CD4+T-cells respond mostly by expanding the effecter/memory pool (321).

My data also suggests that naïve T-cell in early RA are prompted towards Th17 development (which is clearly observed in memory cells at DNA methylation levels on the *IL-17A* gene (322)). The differentiation of Th17-cells *in vivo* remains unclear, while *in vitro,* it can be induced by variable combinations of the pro-inflammatory cytokines IL1-β, IL21, IL23, IL6 with/without TGF-β (127, 323-326). My analysis showed that I*L6/IL21/IL21R* and *TGF- β -1* gene (and the overall gene family of *TGFβ*) were DM in naïve T-cells. The *RORC* gene encoding for RORγt, the master regulator of the Th17 cell lineage itself, was also DM in naïve cells (>5% Δβ-values at 2 CpG sites). Th2 differentiation appears intact in RA (327). In contrast, Th1 polarisation was shown to be compromised by a deficit in Tbet engagement in established RA (96, 97). Interestingly, no evidence of DM were observed on genes of the Th1 or Th2 differentiation cascades (while observed in the Th17 cascade) which add context to the observed difference in the Th17 axis. Defective Th1 polarisation in early RA could therefore be a mechanism resulting in Th17-cells developing preferentially.

An IFN signalling gene node was also highlighted in my String network. Dysregulation of IFNs are often observed in autoimmunity (Systemic Lupus, Systemic Sclerosis, Sjogren's Syndrome, Multiple sclerosis, Dermatomyositis,

Diabetes Mellitus type1, Psoriatic Arthritis) (133, 328-330) suggesting that an activated IFN-response gene expression profile is a common characteristic of chronic inflammatory diseases. In a study of at-risk individual for RA, genes specifically induced by type-1 IFN were indicative of the progression to the inflammatory arthritis stage (331). IFN signatures were so far only associated with outcome in (very) early RA and no longer predictive later in the disease course. As further supported by our data this suggests that IFN-signalling is associated with early pathogenesis, independently of whether this can be exploited clinically later in the disease course. Furthermore, links between IL6 signalling/production and type-I IFN-gene signatures (and vice versa) were also observed in other inflammatory diseases (332, 333), supporting a possible link in early RA development.

Using several techniques, I also confirmed DM of the *TNF* gene promoter in CD4+T-cells of early RA, as well as higher protein levels of several cytokines. String network also shows many other DM gene in *TNF* family and its receptors. The central role of TNF on RA pathophysiology has long been established while my data provide further evidence at a new levels (i.e. DNA methylation) very early in the disease process as well as clearly establishing that this is uniquely associated with naïve CD4+T-cells. TNF has a pro-inflammatory influence on a wide variety of function. It activates leukocyte, endothelial cell, and SFs, inducing the production of cytokines, chemokines, adhesion molecules, and matrix enzymes, activation of osteoclasts and promote angiogenesis (3). TNF can both act on T-cells and can be produced by T-cells (334). TNF can induce the activation and proliferation of naïve and effector T-cells, can also promote apoptosis of highly activated effector T cells, and block Treg suppressive effect (334). Although TNF-α is produced mainly by monocytes, interestingly, my data shows no *TNF* DM in early RA on monocytes, and as well as memory T-cells. This might suggest that in the early events, DM of *TNF* in naïve cells suggest that they are the relevant cells in RA pathogenesis before other cell types. This help supports the central role of naïve CD4+T-cells in the initial stage of RA pathogenesis.

Apart from the genes that were previously known to associate with RA or related to an inflammatory partway, my analysis also brings up other DM genes that were so far unknown to be related to RA, which, may have an important role in RA pathogenesis. This includes genes involved in epigenetic modification itself, DNA methyltransferase (*DNMT3B, SALL3*), Histone deacetylase family (*HDAC2, HDAC4*), and Histone methyltransferase (*EHMT2, PRDM14*). The *DNMT* gene

family is responsible for both de novo DNA methylation or the maintenance of methylated DNA. It plays a critical role during CD4+T-cells development and differentiation where both DNMT and methylation are essential for the appropriate expression of specific genes that help to define lineage and dictate function of T cells (335).

Similarly, Histone modification is also crucially involved in the expression of genes associated with the development of effector and memory T cells (336). Therefore abnormal methylation of these genes involved in the epigenetic machinery in naïve T-cells could lead to defective T-cells capabilities later which may change the ability of T-cell to responds to stimuli and differentiated into particular effector cells.

There are several other new genes that showed up on the string network linked to a known pathway not yet involved in RA. For example, Genes that regulated cell growth and differentiation within the TGFB/SKI signalling cascade; The ubiquitin system and the proteosome encoded gene (*PSMB8, S20 subunit*) linking the *TNF, STAT2, IFN* group of genes on string network to an essential function, the immunoproteasome which processes MHC peptides.

DNA methylation may contribute to pathogenesis via regulating gene activity. Analysis of the downstream process; gene expression and protein transcription, can help confirm the effect of DNA methylation on the ongoing pathology. To validate my list of DM gene, gene expression microarray data from early drug naïve RA patients was retrieved from publicly available sources and compared to my DM list. Many DM genes were matched with a DEG at mRNA level. Serum Cytokine level information retrieved from my supervisor previous work and Elisa experiments on a similar sample set also showed changes of expression for several DM genes. In terms of technical and biological validation, it would have been more informative to study DNA methylation, mRNA transcription, and protein translation in the same patients or with a single cell transcriptome and methylome analysis, which has now gained more interest and is more affordable (notably compared to when my data were acquired). However, utilising publicly available data also remains a useful option where sample resource are limited, while also allowing the use of larger sample numbers. The compatibility of the patient cohorts needs to be considered with caution (both for demographic and clinical parameters) while the technology/platform used and the sample types needs to be adapted to each type of analysis. I took particular care when selecting patients from the tissue banks in order to match their characteristics as much as could be achieved. No statistical difference was noted between the groups

(methylation array, gene expression, ELISA, *TNF* sequencing) except for a slightly longer symptom duration for the patients used in the array due to the choice of 3 swollen joint as a criteria.

After the validation of the possible effect of DM genes at mRNA and protein levels and the interest raised toward a particular pathway (IL6 signalling), a functional study at the cellular level to help decipher the actual contribution of naïve CD4+T-cells to RA pathology would be important. What assay to select depends on that function of the genes/pathways involved. However, this functional study was not within the scope of my PhD. My project will nonetheless allow to design new hypothesis to be tested in a functional study of the associated biological function altered by DM, for which my supervisor is applying for funding.

Within the publication of this part of my PhD, additional work was nonetheless performed in collaboration with other member of the group included as co-authors. A model of unusual T-cell differentiation in RA was proposed by my supervisor group in 2002 (92). My work and the additional work performed in collaboration with other member of the group allow us to propose an updated version of the original model of T-cell differentiation defect in RA (284).

In 2002, my supervisor group reported the development of an atypical subset of naïve CD4+T-cells (CD45RA+ but CD45RO-/dull), in direct relation with levels of in vivo inflammation (measure by CRP), proliferating (1 or 2 cycles, evidence by TREC dilution) but remaining naïve with respect to antigen stimulation while becoming hyper-responsive to TCR (signal 1) + co-stimulation (signal 2) or mitogen (PHA) stimulation, having lost CD62L-. These were hypothesised to results from differentiation following exposure to IL6 (amplified by IL2/TNF) (337), as reported by others (reviewed in (29)). These also had important clinical significance as biomarker for (I) classification from an early arthritis clinic (ii) in relation to the progression of RA from pre-clinical and early inflammatory stages and (iii) for the prediction of 1st line treatment induced clinical remission as well as the stability of remission once achieved (48, 94, 113, 114).

Considering that methylation at a particular CpG position is a binary event (methylated/unmethylated), methylation levels in the form of β-value represent a proportion of cells in the sample that is methylated at that CpG. The differences in β-values between RA and HC observed on the array therefore suggests the emergence of a subpopulation of cells that have altered their methylation status at such positions. Our group therefore hypothesised that the methylation changes affecting cell surface molecules could allow the identification of the subset of naïve CD4+T-cells in which such modifications had happened.

I performed an analysis of the keyword associated with DM-genes and produced a list of potential cell surface marker (Table 4-7). From this list, markers were selected on the basis of their relevance.

An additional experiment was performed to analyse DM cell surface molecule on peripheral blood of 10 HC and 35 early, drug naïve RA patients (Table 4-8, within the same parameters as previous patient groups) using flow cytometry. This experiment was performed by members of a group (co-authors of the paper) (detailed method in Appendix 7). The expression of cell surface molecule (listed below) was first analysed in naïve CD4+T-cells (identified using CD45RA+ and CD45RO-, Figure 4-18, red square).

- *CD4*, showing DM with a Δβ-value of 9.92% ($p=1.80 \times 10^{-4}$)
- *CD62L*, showing low DM with a Δβ-value of 1.26% ($p=1.72 \times 10^{-2}$) but directly down-regulated by IL6 at the gene expression level (337)
- *IL6R*, showing DM with a Δβ-value of 17.05% ($p=3.80 \times 10^{-6}$), or 3.9% ($p=1.30 \times 10^{-5}$)
- *IL2R*, showing DM with a Δβ-value of 21.47% ($p=4.38 \times 10^{-5}$)
- *CXCR4*, showing DM with a Δβ-value of 13.23% ($p=3.45 \times 10^{-3}$)

Table 4-7 DM cell surface molecules

Cluster of differentiation	Cytokine receptors		Chemokine receptor	HLA related markers	IFN related signalling	Others
CD4	*TGFBR1*	*IL15RA*	*CXCR4*	*HLA-E*	*IFNAR1*	*SELL/CD62*
CD27	*TGFBR2*	*IL17RA*	*CXCR5*	*HLA-J*	*IFNAR2*	
CD160	*TGFBR3*	*IL17RC*		*HLA-DOA*	*IFNGR1*	
CD68	*TGFBR3L*	*IL17REL*			*IFNGR2*	
CD300A	*TNFRSF10B*	*IL1R2*				
	TNFRSF10C	*IL1RAPL1*				
	TNFRSF13B	*IL1RN*				
	TNFRSF18	*IL20RA*				
	TNFRSF19	*IL20RB*				
	TNFRSF1A	*IL21R*				
	TNFRSF1B	*IL27RA*				
	TNFRSF21	*IL2RA*				
	TNFRSF6B	*IL2RB*				
	TNFRSF8	*IL4R*				
	TNFRSF9	*IL6R*				
	IL12RB1	*IL10RA*				

Note; DM form naïve String DM gene list (DM-CpG-clusters (score ≥3), isolated-DM-CpG/genes (p≤0.0001), and some manually add gene if score ≥2 or p≤0.001.

Table 4-8 Demographic and clinical data for the control and RA patients used in characterisation of a subpopulation of cells using flow cytometry.

Cohort 3 : Flow cytometry	**HC (n=10)**	**RA (n=35)**
age (years)*	44 (26-59)	54 (26-76)
M/F	5/5	7/13
ACPA (Pos/Neg)	na	15/5
Duration (months)*	na	4 (1.5-12)
TJC	na	9 (0-28)
SJC	na	5 (0-20)
CRP	na	6 (<5-151)

Data are presented as the median (range)

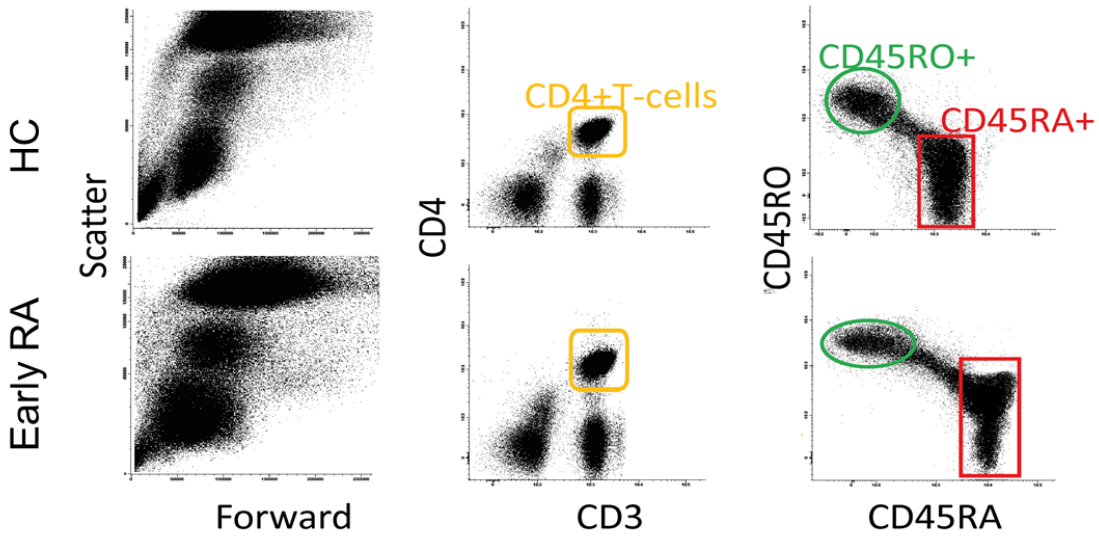


Figure 4-18 The expression of cell surface molecule imported form our publication entitle Differential CpG DNA methylation in peripheral naïve CD4+ T-cells in early rheumatoid arthritis patients, 2020 (284)

Flow cytometry characterised naïve CD4+T-cell of HC and early RA patients. CD3+CD4+T-cells (orange gate) were first gated. Naïve cells were then gated as CD45RA+/CD45RO- (red square) and memory cells as CD45RA-/CD45RO+ (green circle) in a representative HC and RA patient.

The mean fluorescence intensity (MFI) of CD4, CXCR4 and IL2R expression was significantly higher in RA ($p<0.0001$) but not IL6R, which expression was very variable compare to HC (Figure 4-19). The expression of CD62L is either positive (Figure 4-20, A, red circle) or negative (blue square). The % of CD62L- naïve CD4+T-cells was significantly higher in RA (median 1.3%, $p<0.0001$) compared to HC (median 0.15%) and particularly raised in 3 patients with high CRP (55, 75 and 178 mg/L). This demonstrated that CD62L could identify 2 subpopulations of naïve T-cell in RA; the classical naive CD62L+ cells and CD62L- naïve T-cells as described by my supervisor in 2002 (92).

The expression of CD4, IL6R, and IL2R was further analysed between CD62+ naïve, CD62L- naïve and memory CD4+T-cells (Figure 4-20, B, CD45RA- and CD45R+, green circle) in the 3 RA patients that showed a clear CD62L- subpopulation. There was (i) no significant difference of CD4 expression between 3 cell types, (ii) clearly a reduction of IL6R expression on CD62L- compared to CD62L+ naive cells as well as on memory cells, and (iii) a dichotomous IL2R expression on CD62L- (with a positive and a negative peak) compared to CD62L+naïve (all negative) while memory cells were mainly positive.

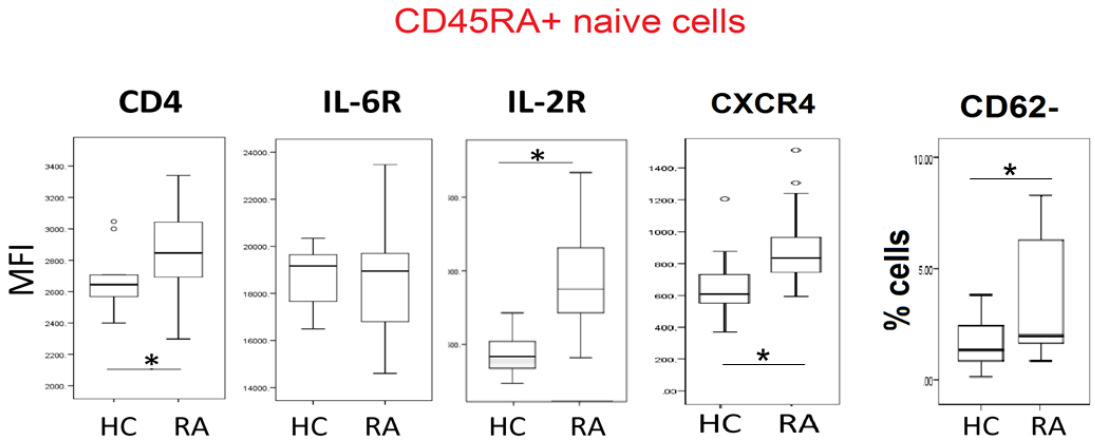


Figure 4-19 Expression of CD4, IL6R, IL2R and CXCR4 imported form our publication entitle Differential CpG DNA methylation in peripheral naïve CD4+ T-cells in early rheumatoid arthritis patients, 2020 (284)

Expression of CD4, IL6R, IL2R and CXCR4 in CD45RA+ naïve T-cells using Mean Fluorescence Intensity (MFI). Results are shown as box plot for 11 HC and 35 RA patients. CD62L was either positive or negative and % of naïve CL62L-cells was recorded and displayed. Significant differences (Mann-Whitney U-test, $p<0.05$) are highlighted by stars.

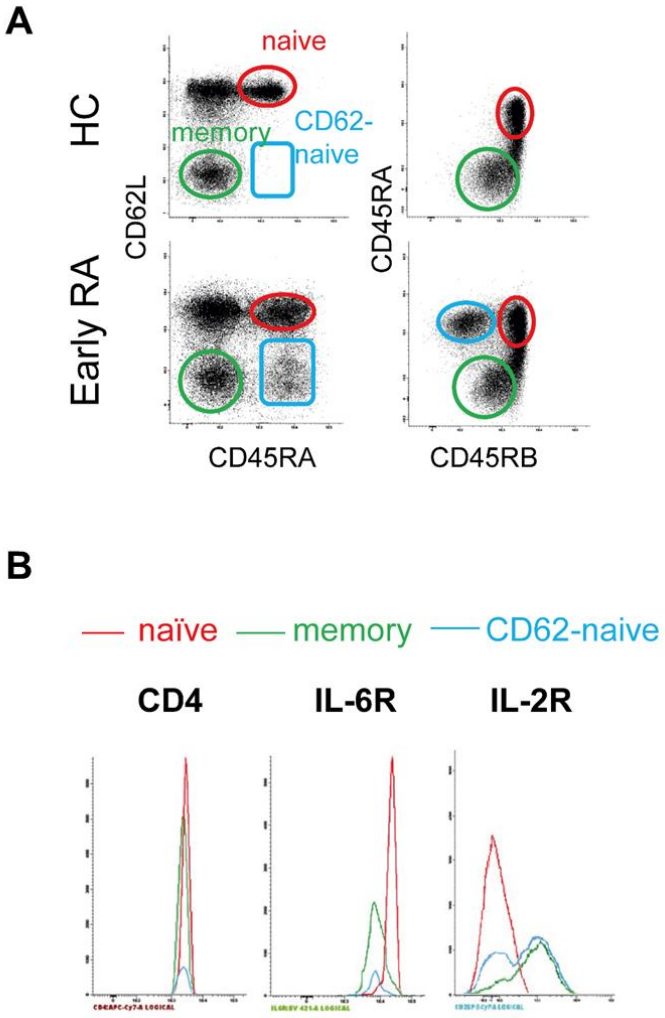


Figure 4-20 Subpopulation of naïve CD4T-cells in RA patients imported form our publication entitle Differential CpG DNA methylation in peripheral naïve CD4+ T-cells in early rheumatoid arthritis patients, 2020 (284)

A) CD45RB and CD62L were further used to refine the phenotype of naïve CD4+T-cells. CD45RB expression was consistently high in naïve cells but declined in experienced cells and was low in memory cells (green circles), with no major difference between HC and RA for this subset. CD62L expression is positive on naive cells (red circle, consistently in HC) but was either positive (red circle) or negative (blue square) in RA defining an subpopulation of naïve CD62L- cells also expressing reduced levels of CD45RB (blue circle).

B) Differential levels of expression for CD4, IL6R and IL2R are shown in a RA patients with a raised CD62L-naïve cells subpopulation (best representative patient displayed) for naïve (red) memory (green) and IRC (blue) cells. Levels of CD4 were not significantly different (n=3). The expression of the IL-6R was lower on CD62L- (MFI 7,300) compared to CD62L+ naive cells (17,600) as well as on memory cells (11,400). The IL2R expression was negative on CD62L+ naïve cells but presented 2 populations (negative <1000 fluorescence units and + fractions > 1000) for CD62L- naïve cells. Memory cells were mainly positive (72% of cells).

Binding to the IL6R by IL6 triggers the formation of a complex with gp130 accompanied by the internalization of the IL-6/IL-6R/gp130-complex (338, 339). Therefore, levels of the IL6R at the cell surface may reflect a balance between recently and past IL6-activated cells (lowered IL6R levels), and non-activate/resting cells (high IL6R levels). Thus, a large distribution of IL6R levels as observed in total naïve cells is not unexpected, if this observation was used to support the hypothesis of a proportion of naïve cells undergoing IL6 driven differentiation. IL6 signalling was shown to directly reduce the expression of CD62L (337). As Such, the naïve cells that have lost CD62L expression, have also reduced their expression of the IL6R (suggesting exposure, signalling and internalisation of the receptor) for most of them, and may therefore directly represent the subset of naïve cells that are undergoing IL6 driven differentiation.

IL6 was reported to increase the expression of *DNMT1* which correlate with DNA methylation in T cells (244, 335), suggesting a direct link between inflammation and alteration in DNA methylation (340). The main conclusion of my work is therefore that IL6 could induce the change in DNA methylation of the *DNMT1* gene (Δβ-value of 8.86% (p=0.0224) as well as several other genes involved in the epigenetic machinery (i.e.*DNMT3, SALL3, HDAC2 , HDAC4, PRDM14*, and *EHMT2*) observed in early RA, resulting in an abnormal activity of these genes leading to abnormal DNA methylation in other genes (as well as other modifications that needs to be explored further as well).

IL6 signalling also alters the chemokine receptor balance at the surface of T-cells, silencing CD62L while increasing chemokine receptor expression for pro-inflammatory factors such as CCL1/MCP-1 or CCL12/SDF1 to allow cells to migrate to inflamed tissue (341). CD62L- naïve T-cells may also exhibit other changes resulting from alterations in DNA methylation, these cells, previously maned Inflammation related cells (IRC) in a publication from my supervisor's group (94), were shown to have specifically increased their expression of CXCR4 (also DM with a Δβ-value of 13.23%, p=3.45 $x10^{-3}$), CXCR5 (Δβ-value of 5.50%, p=4.07 $x10^{-5}$), CCR3 (Δβ-value of 19.54%, p=8.73 $x10^{-4}$). In addition, another recently described subpopulation of CD4+T-cell, Follicular T-cell (TFh), are also indirectly induced by IL6, via IL21 (342). These TFh cells display CXCR5+/$PD1^{high}$ phenotype (343, 344). They have recently been observed in RA (345). Both these markers were modestly DM (*CXCR5* Δβ-value of 5.50%, p=4.07 $x10^{-5}$ and *PD1* Δβ-value of 6.69%, p=6.05 $x10^{-3}$) while considering the IL21 ELISA results, and the high expression of CXCR5 expression previously reported on CD62L- naïve cells in RA patients (94), these data altogether further support the potential of

IL6/IL21 in generating CXCR5+ cells that may correspond to (i) CD62L- naïve cells, and/or (ii) TFh.

The new data present in my thesis, enabled me to redefine the original model of the IL6 driven alteration in naive CD4+T-cells presented in 2002 by my supervisor (92) which proposed the perturbation of CD4+T-cells differentiation in early RA, as a result of IL-6 activation of naïve T-cells, introducing alteration of DNA methylation driven by IL-6 as a possible mechanism for the generation of the unusual population of CD62L – naïve CD4+T-cell (92, 94, 243, 337). DM in genes enables changes in the DNA accessibility to several genes in pathways such several signalling cascades (TNF, TGF, IL2/15/21) , Interferon signalling related genes expression, the polarisation of Th17 cells. The effects of these changes together with functional alteration in naïve CD62L- CD4+T-cells may then allow them to migrate to the joints and contribute to the development of chronicity via the acquisition of resistance to apoptosis as previously suggested (137, 346) and the local maturation of Th17 cells. In Figure 4-21, I therefore propose an updated version of the original model of T-cell differentiation defect in RA.

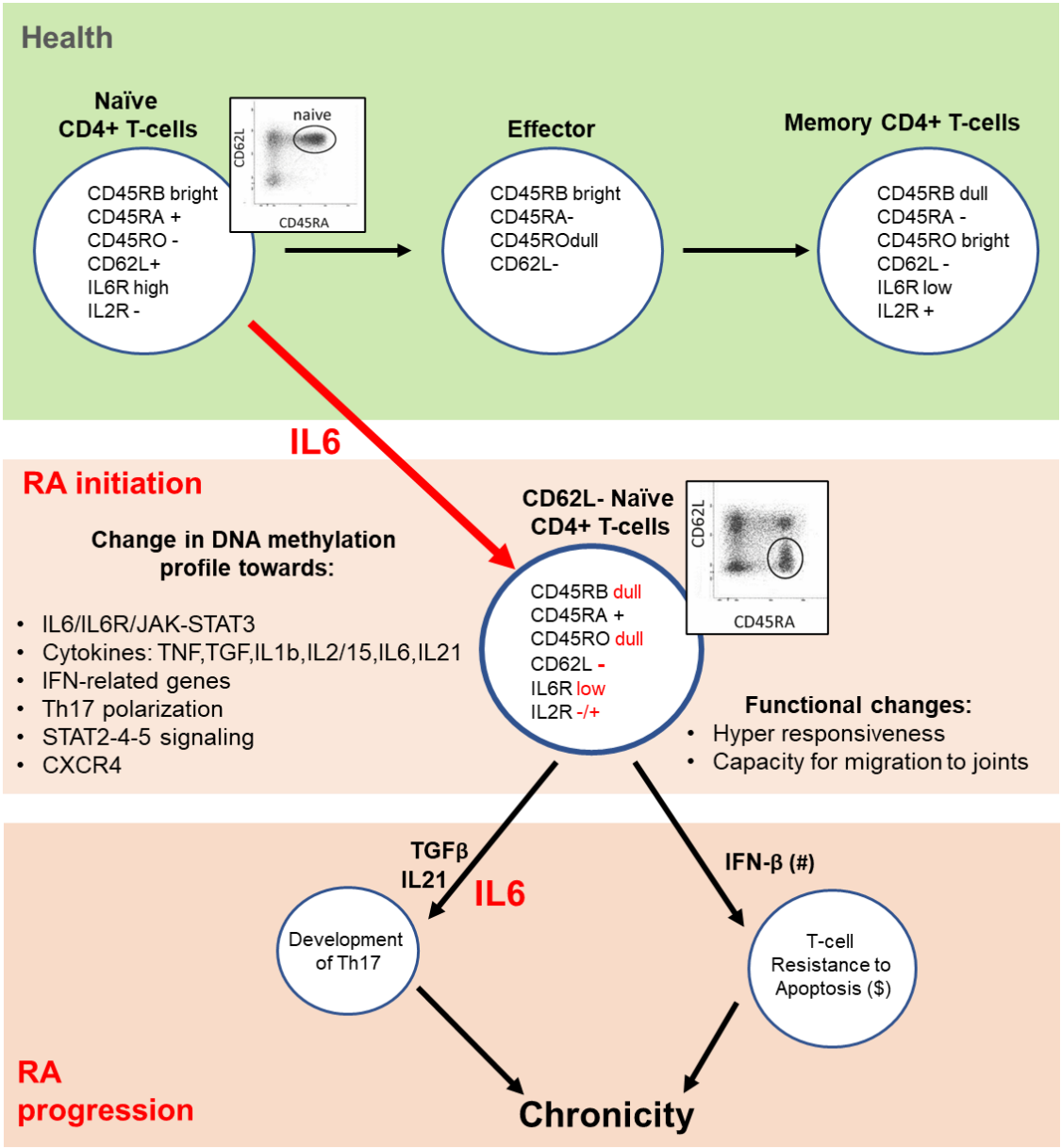

Health
Naïve
CD4+ T-cells
naive
CD62L
CD45RA
CD45RB bright
CD45RA +
CD45RO -
CD62L+
IL6R high
IL2R -
Effector
CD45RB bright
CD45RA-
CD45ROdull
CD62L-
Memory CD4+ T-cells
CD45RB dull
CD45RA -
CD45RO bright
CD62L -
IL6R low
IL2R +
IL6
RA initiation
Change in DNA methylation profile towards:
IL6/IL6R/JAK-STAT3
Cytokines: TNF,TGF,IL1b,IL2/15,IL6,IL21
IFN-related genes
Th17 polarization
STAT2-4-5 signaling
CXCR4
CD62L- Naïve
CD4+ T-cells
CD62L
CD45RA
CD45RB dull
CD45RA +
CD45RO dull
CD62L -
IL6R low
IL2R -/+
Functional changes:
Hyper responsiveness
Capacity for migration to joints
TGFβ
IL21
IL6
IFN-β (#)
Development of Th17
T-cell Resistance to Apoptosis ($)
RA progression
Chronicity

Figure 4-21 Hypothetical model of how IL6 and naïve CD4+T-cell may contribute to the development of chronicity. Modified from our publication entitle Differential CpG DNA methylation in peripheral naïve CD4+ T-cells in early rheumatoid arthritis patients, 2020 (284). In health, naïve CD4+T-cells differentiate into effector cells and memory cells as displayed in the green box. Cell surface markers used to identify CD4+T-cells at different stages are displayed: naïve ($CD45RB^{bright}CD45RA^{+}CD45RO^{-}CD62L^{+}$), effector cells, and memory cells. An atypical subset of naïve CD4+T-cells ($CD45RB^{dull}CD45RA^{+}CD45RO^{dull}CD62L^{-}$) is observed in early RA (light orange box)(92) sharing both characteristics of naïve (maintaining $CD45RB^{high}$, $CD45RA^{+}$) and memory (CD62L- and $CD45RO^{dull/+}$) CD4+T-cells. In addition to the loss of expression for the lymph-node homing receptor (CD62L), other cytokine receptor expression was modified (e.g. $IL6R^{low}$ and $IL2R^{+/-}$), as shown during my PhD. Previous work (92, 337, 347) suggesting that this atypical differentiation of naive T-cells in RA was possibly induced by IL6. It was also associated with functional changes such as hyper-responsiveness (92) and capacity for migration (expression of CXCR4 (94) confirmed during my PhD). My work shows that DNA methylation is a possible mechanism underlining the changes observed (as depicted in the dark orange box). It is reported that IL6 can induce change in epigenetic machinery gene (*DNMT*) (244, 335) which can affect the methylation profile of many genes/pathways as highlighted in the gene network shown on Figure 4-15. IL6 was also showed to directly reduce expression of CD62L to change the migration pattern of cells towards IL6 expressing tissue (341). The cells undergoing these changes have acquired the capacity to move to tissues, display DM in genes related to type-I IFN-signalling and TNF-α signalling as well as being prompted to differentiate towards Th17 subset. Upon, migrating to the joint (CCL12/SDF1 attracting them via CXCR4 for example (94), while being CD62L-), they can encounter further signal (IL6 notably expressed by synovial fibroblasts) which will complete their transformation towards Th17 cells for example, while INF-β (also expressed by synovial fibroblasts (137) can provide both a survival signal as well as further inflammatory signal.

Chapter 5 Results Part2 Biomarker development

5.1 Introduction

The second goal of my thesis was to develop a DNA methylation biomarker for RA diagnosis. In this chapter, I introduced the biomarker of RA diagnosis currently used, the need of new biomarker, the method to detect DNA methylation, and the basic terminology used in biomarker research as well as some statistical analysis that related to biomarker development.

5.1.1 Biomarker in RA

RA is the most common inflammatory arthritis involving immune cells activation, cytokine, reactive oxygen species (ROS), and protease production. Leaving the condition untreated or receiving an ineffective treatment, leads to the destruction of joints, loss of physical function, and wild spread of inflammation to the other part of the body. It is widely accepted in the rheumatology community that aggressive/effective treatment of RA in the early stages of disease offer a chance to control the disease in the long term when it is still susceptible to treatment, also known as the window of opportunity (348, 349), and importantly prior to the development of irreversible damages. Early diagnosis and early access to effective treatment are keys to prevent the irreversible damage and improve disease course of RA patients while the evidence of the benefit of early treatment have been established (350-352)

Early arthritis clinic (EAC) and inflammatory arthritis outcome

Currently, patients who have pain, swelling and show signs of inflammation in their joints are sent to an EAC where an early inflammatory arthritis (IA) is diagnosed, treated and monitored. RA is the most common inflammatory outcome observed in EAC. The characteristics of RA were described in the general introduction. Other IA commonly diagnosed are psoriatic arthritis, reactive arthritis, gout or undifferentiated arthritis (where the symptoms remain unclear and progressed very slowly), while a proportion of people also do not develop persistent symptoms and return to normal. However, at presentation all IA share the same features of inflammation such as pain, swelling, stiffness, redness, warm joints, affecting one or more articulations. Some patients shows specific features such as skin symptoms suggesting psoriatic arthritis or have

specific autoantibodies helping to differentiate from RA notably the antinuclear antibody (ANA) for systemic lupus erythematosus (SLE) and other connective tissue diseases for example. However, not all patient express those specific features so there is always cases where it is difficult to establish a diagnosis. Furthermore, IA patients who develop RA are not all positive for RA features (notably ACPA or RF) rendering the overall classification of Ab-negative patients difficult.

Other inflammatory arthritis

Psoriatic arthritis is an asymmetric polyarthritis. This often associated with psoriasis symptoms. This arthritis is sometimes difficult to distinguish from RA as it has no specific autoantibodies marker and the joint symptoms of psoriatic arthritis may precede the onset of skin symptoms by many years. Evaluation and monitoring for other signs such as nail changes or sausage toe, spinal involvement, family history can help with the diagnosis.

Reactive arthritis often presents as a monoarthritis in large joints, such as the knees ankles and feet and also involved the tenosynovium, entheses, and surroundings. Reactive arthritis is triggered by an infection in another part of the body (i.e. intestines, genitals or urinary tract). Unlike other types of IA, reactive arthritis lasts a relatively short time, usually around three months to a year, however, it can last longer in some patients and can have random flare-ups years after first symptoms.

Undifferentiated arthritis (UA) is an inflammatory oligoarthritis or polyarthritis that cannot be classified in the early stage. Over time it may turn into remission spontaneously in about ~30% of patients or evolve into a chronic inflammatory disease, even sometimes progressing to RA with time(1).

RA Classification

RA should be suspected in the patient who presents with inflammatory polyarthritis. Medical history acquisition, physical examination, along with selected laboratory testing to identify features that are characteristic of RA or that suggest an alternative diagnosis is required for the initial evaluation of such patients.

The new RA classification criteria proposed by the American College of Rheumatology and the European League Against Rheumatism developed

(ACR/EULAR) in 2010 (353) improve clarification form the old 1987 classification criteria, allowing patients to be classified at earlier stages. In this criteria set, RA is defined based on the presence of synovitis in at least one joint that cannot be explained by the other diagnosis, and achievement a total score of ≥6 out of 10 points. The score comes from 4 domains: Joint involvement (range 0–5), serological abnormality (range 0–3), Acute-phase reactant (range 0–1), and symptom duration (range 0–1) (353).

- **Joint involvement**
  The number and site (at small > large joints) with swelling or tenderness on examination that is indicative of active synovitis use in scoring. The higher number of joint involvement (especially the small joint) added up more point which can contribute to the score up to 5.
- **Serology marker**
  The abnormality of two sera marker, RF and ACPA are included in this criteria.
  - RF is an auto-antibody defined as the anti-bodies against Fc portion of IgG. The predominate RF antibody is IgM but it can also be in the form of any isotype of immunoglobulins, i.e. IgA, IgG, IgM, IgE, IgD.
  - ACPA is an auto-antibodies against citrullinated peptides which is the result of posttranslational modification of arginine by the enzyme peptidyl arginine deiminase (PAD) in response to inflammation, apoptosis or keratinization (354).
  - A positive test in either of RF or APCA markers contribute to the a score by 2- 3 points.
- **Acute-phase reactant**
  The elevated acute-phase response, either from C-reactive protein (CRP) or Erythrocyte sedimentation rate (ESR) contribute to the score by 1 point. Both are a marker for inflammation.
  - CPR primarily produced by a liver in response to increased levels of inflammatory cytokines, especially IL6. The level of CPR in plasma increase at least 25 times during inflammatory conditions. The level of CPR increase or decrease rapidly upon the presence or absence of stimuli. It is used as a clinical marker of acute phase inflammation and also a predictor of cardiovascular disease(355).
  - ESR is the rate at which erythrocyte in WB sample settles at the bottom of a test tube over a period of one hour. A higher rate than normal indicates inflammation. ESR is a result of the balance

between pro-sedimentation factors, mainly fibrinogen, and those factors resisting sedimentation, the negative charge of the erythrocytes. The higher proportion of fibrinogen which response to an inflammation cause erythrocyte stick to each other and sedimented quicker.

- **Duration of symptoms**

  Duration of symptoms is the maximum duration of signs or symptoms of pain, swelling, stiffness of any joint that is clinically involved at the time of assessment self-report by patients. The symptom of duration ≥ 6 weeks contribute to a score of 1.

Around 70% of IA patient who develop RA can be diagnosed by the 2010 criteria (sensitivity ~80-85%). There are still many people who experience a delay in diagnosis especially patient with no autoAb (i.e. no RF/ACPA). RF is present in 70-85% of people with RA but also in many other conditions (hence low specificity ~ 40% (356, 357)). Although the negative result of RF or APCA does not exclude patient from an RA diagnosis, it slows down the process. People have to wait for more joint to become involved to meet the RA classification criteria and therefore have delayed access to treatment. ACPA is an important marker to help with the classification of RA with high specificity of (95-98%) (356) while ACPA positivity was found in ~50-60% of patient with RA (356, 358, 359).

Treatment of RA / IA

The goal of the medication treatment for RA and other IA is to reduce the inflammation, relieve pain, and prevent or slow down disease progression that can cause joint damage and loss of function as well as other systemic manifestations. Medication treatment for early RA and other arthritis including Disease-modifying antirheumatic drugs (DMARDs), Non-steroidal anti-inflammatory drugs (NSAIDs), and steroid.

DMARDs are a class of drugs that work as immunosuppressive and immunomodulatory agents(360). It was indicated to relieve pain, reduce inflammation and slow the progression of the disease. It was recommended by the 2016 EULAR recommendations for the management of early arthritis to used DMARDs as early as possible, preferably before the onset of erosions to reduce and prevent the risk of further joint damage(361). There are two classed of DMARDs, conventional synthetic DMARDs (cs-DMARDs) and biologic DMARDs

(b-DMARDs). cs-DMARDs broadly act as mild immunosuppressor (inhibit/reduced cell proliferation). The commonly used drugs such as MTX, leflunomide, hydroxychloroquine, and sulfasalazine. The b-DMARDs target a specific pathway of the inflammatory cascades on immune cells signalling. For example anti-TNF agents (adalimumab and infliximab), IL-6 inhibitor (tocilizumab), and Jak inhibitors (tofacitinib) or anti-costimulation (CTLA4-Ig). cs-DMARDs have been used for over 2 decades and the side effects are already well known. It is available orally and inexpensive. As cs-DMARDs is a slow-acting drug, it takes several weeks to initiated effects, NSAIDs or/and corticosteroids is given to patients to relieve pain and reduce inflammation while waiting for the effect of DMARDS or in case of flare stage (disease active stage). The response to the treatment is different from patients to patients. It needs close monitoring to observe the patients response. Drugs adjustment or change to more appropriate treatment is sometimes necessary.

Increasing evidences show that b-DMARDs have consistently better efficacy for both early and established disease (362), however, it only available via injection or infusion and also pricey (£5-25k/years). Thus this prescribed with restriction (NICE guidelines) is only available for active RA characterised by a high disease activity index (DAS28 >5.1), and disease resistant to cs-DMARD.

According to the latest recommendations of EULAR and the American College of Rheumatology(159), MTX is still the first-line drug for RA treatment.

Clinical parameters to monitor the progression of the disease or the response to treatment

The parameters used (notably in my project) to monitor disease progression or for the evaluation of the response to treatment are included but not limited to:

Joint count

Tender joint count (TJC) and swollen joint count (SJC) are the examination of an individual joint for signs of pain (for TJC) and swelling (for SJC). The number of affected joints are counted. The total number of joints that will be examined are various depend on the particular assessment. It could go up to 68 joints on some scales but in the most standard system in a patient care unit and research trials is the 28-joint count which includes shoulders (2), elbows (2), wrists (2), knees (2), proximal interphalangeal joints (10 joints), and metacarpophalangeal joints (10).

DAS-28

Disease Activity Score (DAS) is a measure of disease activity in RA. DAS-28, Disease activity score examined in 28 joints is the most common used. The score is calculated based on TJC, SJC, the acute phase reactant (CRP or ESR), and the patient global assessment of health. A DAS28 of greater than 5.1 corresponds to high active disease, between 5.1 to 2.6 correspond to moderate and low disease activity, and less than 2.6 implied the remission.

5.1.2 Epigenetic Biomarker

Epigenetic pattern can be altered in response to the environment and therefore are reversible. Such dynamic changes provide useful information on a specific condition at a specific time and may have use for medical management. The use of epigenetic marks as biomarker therefore gained more attention and are already being used in the clinic for some disease reviewed in (363).

An epigenetic biomarker can be defined as “any epigenetic mark or altered epigenetic mechanism that can be measured and evaluated as an indicator of biologic process, pathogenic process, or pharmacologic response to a therapeutic intervention. Epigenetic biomarkers are of interested in many diseases such as cancer, psychiatric and neurodegenerative disorders, and chronic inflammatory diseases including RA (364-367).

A lot of work has been done in the cancer field. At the moment histone modifications at specific sites are still technically much more difficult to measure than DNA methylation. It is also not clear how stable these histone modifications are. DNA methylation marks are therefore the most commonly used among other types of epigenetic change.

Several DNA methylation biomarkers are already used clinically for cancer early detection, diagnosis, treatment monitoring and prediction of response, and prognosis (Table 5-1 and More candidate genes that have been approved for the clinical use and the candidate genes with a potential in the discovery research could be found in the reviews of Warwick J. Locke and Yunbao Pan(368, 369)).

In the RA field, the study of the epigenetic landscape is growing. Most of the work is in the discovery phase that still searches for differentially methylated genes between RA and heath or other IA diseases to identify promising biomarkers candidates.

Table 5-1 DNA methylation biomarkers are used clinically for cancer.

Disease and purpose	Gene / Epigenetic mod	References
Screening for Prostate cancer	The measurement of *GSTP1* or *APC* promoter methylation in plasma, serum, or urine samples.	(370, 371)
Biomarker for early clinical stage Colorectal cancer.	DNA methylation of *SEPT9* detectable in the plasma is able to distinguish colorectal cancer and adenomas from normal and inflammatory colonic tissue.	(372)
Diagnostic for Lung cancer and prognostic patients suffering from malignancy	*SHOX2* methylation. Higher levels of *SHOX2* methylation in pleural effusion samples demonstrated a shorter overall survival.	(373)
Predict the response to chemotherapy for Glioblastoma	Methylation of the promotor of *MGMT* in blood or tumor biopsy is used to predict the response to alkylating chemotherapy such as Temozolomide	(374)

Technique to study of DNA methylation at specific CpG/region

Identified DM candidate CpG sites/genes with potential as a biomarker need to pass through a series of steps including verification of the methylation status, validation of methylation pattern in a large cohort successfully to establish their value as biomarker in a development process.

DNA methylation is a modification of cytosine that occurs only at CpG dinucleotide sites but that does not change the base itself. Several techniques can be used to study DNA methylation at a specific site, the main one being

- bisulfite conversion followed by downstream methods such as
    - PCR and sequencing
    - Pyrosequencing
    - Methylation-Specific PCR (MSP)
    - quantitative Methylation-Specific PCR (qMSP)
    - PCR with high resolution Melting.

In my thesis, the techniques that were selected for verification of DNA methylation status of the target sites were bisulfite sequencing, followed by the development as a biomarker assay using qMSP.

5.1.3 Biomarker Development

The goal of biomarker research is to discover and validate assays that can be used in a clinical settings. Biomarkers could be used to gain information of the presence or absence of disease (diagnostic biomarker), the patient prognosis (prognostic biomarker), the response to a specific intervention (predictive bio-marker), the effects of ongoing treatment (therapy monitoring biomarkers), or future risk of disease development (risk markers)(375).

In general, biomarker development begins with the discovery of a “mark” or “events” that changes with respect to a particular situation. After the identification of such a “mark” the changes with respect to this situation need verification and then secondary validation in order to establish it as a potential candidate. To see if the biomarker offers potential utility, it needs to meet acceptability of performance characteristics both technically and with respect to its performance as an outcome indicator. The assay once validated (again in an additional large number of samples) may be used in clinic prospectively before receiving an approval to use be used on a wider market (and/or be commercialised).

Characteristics of biomarkers

Important universal characteristics for biomarker are non-invasiveness, easy to measure, robust, and inexpensive. The sample should ideally be taken from readily available sources, such as blood or urine. It should be the less process samples (for example, using the WB or PBMC instead of specific cell subset that needs more steps for isolation or purification). The technology chosen for the assay should be easy to perform by routine clinical services (hence PCR over sequencing for example), give reliable result, and have low inter-user variability.

The most important characteristic of the biomarker remains its fitness to propose. For example, a diagnostic biomarker should have high sensitivity (meaning being present in a high proportion of the patients with the disease) and specificity (being present ONLY in that disease) for the disease of interest. It should be able to detect the small differences if quantitative, notably in the early stages of the disease where it may not (yet) be increased/reduced as much as in established condition, but importantly it also should be able to discriminate between the disease of interest from people with similar symptoms but with a different disease while it may have been identified compared to healthy people. All of these suggest that the best biomarkers may be those underlying a disease mechanism however, it is not mandatory (376).

To put in the content of my thesis, I aim to develop a DNA methylation biomarker assay for the classification of RA. This biomarker should be specific, discriminating RA patients form those with other inflammatory arthritis (IA) diseases. I therefore focused on DM CpG candidates in CD4+T-cells, which are key to RA pathogenesis and involved in early processes. The sample material for the assay would be easily accessible (hence blood, rather than tissue biopsies for example), while, whole blood (WB) or PBMC are more convenient than purified CD4+T-cells for translation into future clinical practice. The technology of choice is a qMSP assay (based on qPCR) which is highly sensitive while practical and already used by hospital services.

5.2 Objectives

For the second part of my PhD, my overall aim is to select potential CpG candidates for the development of a biomarker assay using the qMSP technic for RA classification.

Objectives are:

- Develop a strategy to select candidate CpG(s)/gene(s)
- Validate/verify the methylation status of candidate CpG from the array data by sequencing
    - In HC 5 cell subsets
    - In HC vs early RA patient using CD4+T-cells
- Choose the best candidate CpGs/genes based on the technical requirement of using qMSP
    - Develop the biomarker assay for the top candidate genes
    - Test the assay in the different sample types (WB, PBMC, CD4 cells)
    - Test the assay on small number of patients attending the Leeds early arthritis clinics (RA or non-RA)
- Test large number of samples to demonstrate the potential of the biomarker assay (notably in autoAb negative patients)
- Build a predictive model to demonstrate the added value of the biomarker
- Replicate the overall work in a second cohort
- Examine whether the assay may also have added value for prediction of response to 1st line therapy in early RA (MTX)

5.3 Result

This chapter describes the whole process of how I have developed an epigenetic DNA methylation biomarker assay for RA classification from selecting the candidate CpG targeted by the assay through to the validation in patient samples.

According to the role of DM identified in T-cell (compared to monocytes) in the early stage of RA presented in chapter 4.3 Part1 result, I decided to focus on methylation change in T-cells in RA patient to select a DM-CpG candidate to develop the biomarker assay.

The technology chosen for the final assay is qMSP, using DNA from a blood samples. This assay is used the quantify the methylation status of the target CpG sites (the marker) by using specifically design primers that would only generate a qPCR product if the target DNA is either methylated or demethylated.

For the assay development, I also needs to consider how the assay can be used in daily practice, which means using the least processed samples (i.e. WB or PBMC) rather than purified CD4+T-cells to limit the number of steps that may introduce bias. qMSP detect total DNA methylation of the target site in the whole DNA sample. The challenge in developing such a biomarker assay is how to detect the methylation change that occurs in T-cells (which are my target cells) while working on the mix population of cells in the PMBC or WB sample. CD4+T-cells are account for 45-70% of PBMC and around 15% of cells in WB. The difference in methylation status in T-cells may end-up being too dilute to be observed because there is also a methylation signal from other cells type in the sample and this is a major restriction to the assay design.

Designing the assay to enable detection of the DM in only the target cell in the presence of several cell types in the sample, requires a strict process for the specific candidate CpG selection as a 1st step together with the general qMSP assay design. The important selection criteria is that the methylation status of the candidate marker (CpG sites) in the target cell is different from the status in other cell types (ie completely demethylated in Target cells, while complete methylated in non-target cells - or vice versa). Specific primers would then only generate a qPCR product, if the target DNA is either methylated (or demethylated) so that only the methylation status of the target cell can be detected. A diagram in Figure 5-1 shows this overall principle.

This principle has been successfully used to develop assays for quantifying 2 types of T-cell subsets, namely Treg and Th17 cells in blood samples, resulting in commercial assay (377-379). In these situations, the *Foxp3* and *IL17* genes

showed specific demethylation in these 2 T-cell subsets respectively, while all other T-cells but also all other lymphocytes and monocytes in the blood, showed full methylation at the chosen candidate CpG site targeted by the assays.

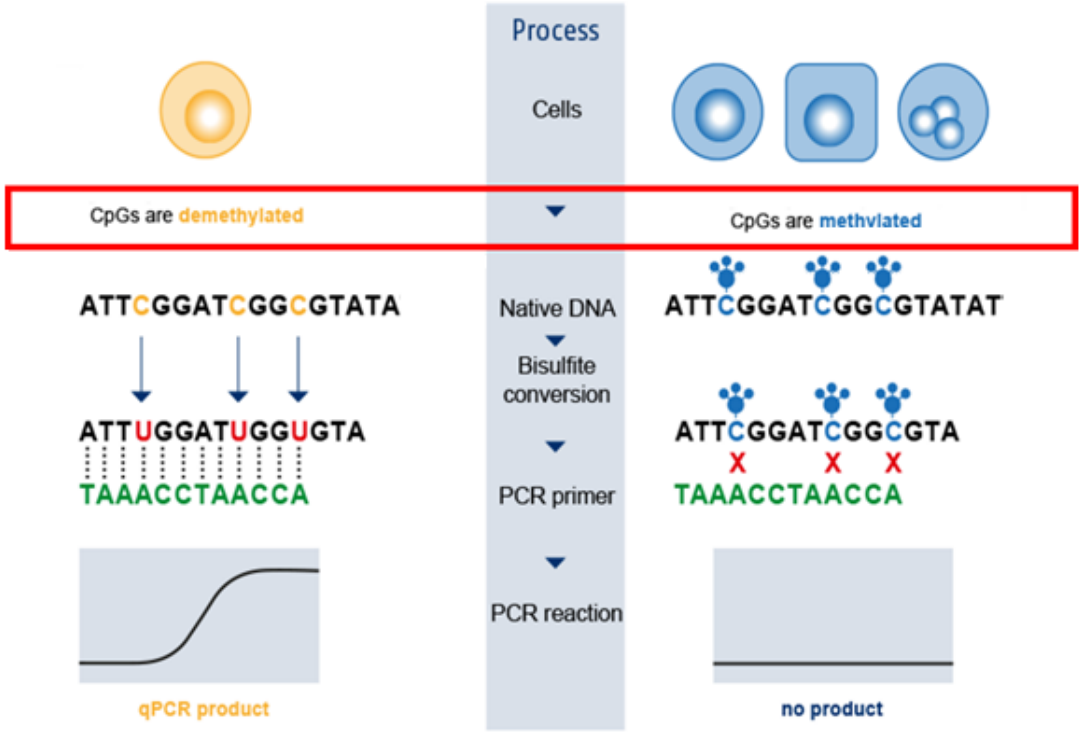


Figure 5-1 Principle of a qMSP assay for the detection of DNA demethylation in a specific cell population (here in yellow), while working with DNA from a mix cell population sample (the blue cells added to the yellow cell). The requirement for this assay is that the CpG candidate is epigenetically active (fully demethylated), exclusively in the target cells while fully methylated in other non-target cells in the sample. In target cells the cytosine will become uracil after bisulfite conversion, while in non-target cells, the CpG marker being methylated, cytosine are protected by the methyl group and will not be converted. Next, the primers are designed specifically to match the concerted sequence with demethylated CpG, it can then bind and amplify a PCR product in the target cells. While in the non-target cell, the same primer cannot bind due to mismatches, thus no PCR product is amplified. This allows the detection of methylation status of the specific cell population as a % of the total mix cell population (quantified with a similar assay on a house keeping gene. Figure adapted from the Epiontis website (377, 379).

The workflow of my biomarker development is illustrated in Figure 5-2. The first step was the identification of the CpG sites to be targeted in the assay (Figure 5-2, A). Genome-wide methylation array data (illumina 450K array) (which contain the methylation information of more than 485,000 CpGs), were used as a resource to find the candidate CpG in CD4+T-cells that may have value as a biomarker to identify the difference of methylation between RA and other clinical groups. Several data sets were used to refined strategic decision to inform the target CpG selection regarding non T-cell methylation status.

- Our dataset (GSE121192) in early RA and HC
- Other publicly available dataset were used including :
    - Different cell types
    - Different disease groups

The second step was the target verification (Figure 5-2, B). After obtaining a list of candidates CpG, the methylation status of these candidate CpG was then verified to establish whether the array data could be replicated by bisulfite sequencing.

The third step is development of the qMSP assay itself, to detect DM of the candidate CpG site in sample (Figure 5-2, C). This included the development of the PCR assay and the optimisation of the best type of samples to use (WB, PBMC or purified CD4T-cells) using a few DNA samples with known diagnostic information (i.e. HC versus RA or UA).

The final step is assay validation in a real cohort of samples (Figure 5-2, D). The value of the assay itself for the classification of patient and the performance of the qMSP assay needs to be established in a relatively large number of samples.

The added value of the assay over the current available biomarker/signs and symptoms (i.e. demographic and clinical parameter) used for RA classification then need to be demonstrated for the assays toward added clinical value.

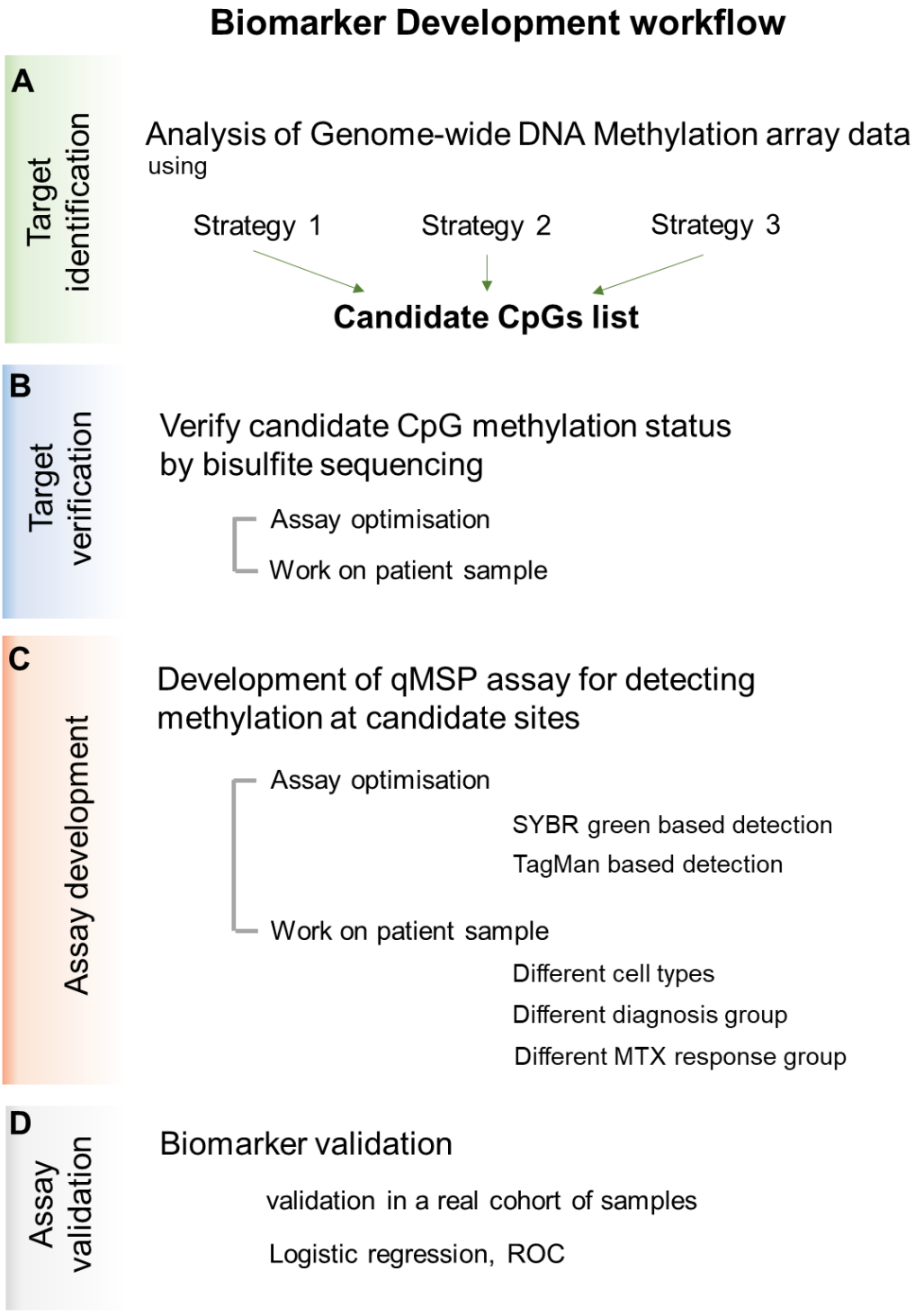


Figure 5-2 Biomarker development workflow. A)Target identification, B) Target verification, C) assay development and D) assay validation.

5.3.1 Selecting candidate CpG Targets : Analysis of 450K DNA methylation dataset

5.3.1.1 Selection strategy 1: CpG candidate from our dataset

This strategy selects candidate CpG according to the qMSP assay principle with demethylation in the target cells, while allowing the assay to work from a mixed cell population sample. The first requirement of this strategy is that the methylation status of candidate CpG in the target cell, (here, CD4+T-cells), is clearly different in all other cell types in the blood. I chose to select CpGs that were demethylated in T-cells (ideally with a β-value ~ 0 to 0.2 in array data) while methylated in the other cells (ideally β-value ~ 0.8 to 1). Therefore, following DNA bisulfite conversion, the DNA sequence of T-cells will be distinct from that of all other non-targeted cells and could be amplified using specifically designed PCR primers for the demethylated sequence, while there will be no amplification from non-target cell DNA. This way, it is possible to detect the methylation signal just in T-cells. However, the limitation here, despite the benefit of having a highly specific T-cells methylation assay, is that if the candidate CpG is DM between HC and RA it may also be shared with other diseases as for example related to an inflammatory cascade.

To select the candidate CpG sites that would allow RA patient to be differentiated from HC, I first analysed the Illumina 450K genome-wide methylation data (the same GSE121192 dataset from the previous chapter Table 5-2) which contain methylation data of 6HC and 10 early RA of naïve CD4+T-cell, memory CD4+T-cell and monocytes. After pre-processing the dataset , quality checked, removing CpGs that were known SNP or removing the probes known to cause cross-reaction, the list of DM CpGs between HC and RA in naïve CD4+T-cells (established as described in chapter 3.4), I selected several candidate CpG sites from over 485,000 CpG by utilising 3 filtering steps.

1. DM between HC and early RA in naïve CD4+T-cells [p-value ≤0.0001, scoring rule1=1] and having at least 1 significant DM GpG nearby [scoring rule3>1].
2. Demethylated DNA in naïve CD4+T-cell of HC sample [β-value 0 to 0.6]
3. Methylation in the monocytes dataset for HC sample [β-value 0.8 to 1] and no DM between HC and RA in monocytes.

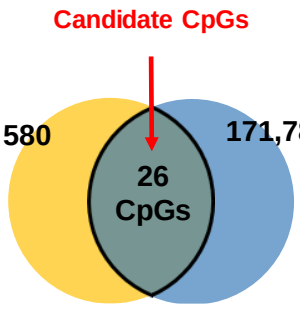
Table 5-2 Illumina 450K dataset used in a different strategy.

Strategy	Dataset ID/GEO accession	Samples	Cell types
Strategy 1 : 1 dataset	GSE121192 (our data)	6 HC, 10 early RA	Naïve CD4 T-cells, Monocytes
Strategy 2: 4 dataset	GSE121192 (our data)	6 HC, 10 early RA	Naïve CD4 T-cells, Monocytes
	American research group	31 HC	Naïve, Monocytes, B-cells
	GSE71841	12 HC	CD4+T-cell
	GSE35069	6 HC	CD4+T-cells, CD8+T-cells, NK cells, B-cells, Monocytes, Neutrophil, Eosinophil, Granolocytes
Strategy 3: 7 dataset	GSE121192 (our data)	6 HC, 10 early RA	Naïve CD4 T-cells, Monocytes
	American research group	31 HC, 63 RA	Naïve, Monocytes, B-cells
	GSE71841	12 HC, 12 RA	CD4+T-cell
	GSE 111942	18 HC, 25 RA	PBMC
	GSE 87095	73 HC, 49 RA	B-cells
	GSE 117929	19 HC, 18 Ssc	PBMC
	GSE 82218	25 HC, 30 SLE	PBMC
	GSE 88824	14 HC, 13 MS	WB

Ssc: Systemic sclerosis, SLE: Systemic lupus erythematosus, MS: Multiple sclerosis

Figure 5-3 Results of the filtering criteria to select candidate CpG by my 1st strategy.

	Target cell condition
Sample	HC vs early RA
Cells Types	naïve CD4+T-cell
HC β-value	≤ 0.6
DM, p-value	≤ 0.0001
DM, rule3 score	≥ 1



Non-target cell condition	
HC vs early RA	**Sample**
monocytes	**Cells Types**
≥ 0.8	**HC β-value**
≥ 0.01	**DM, p-value**

The filtering results workflow is shown in Figure 5-3. Ideally, the β-value were set at 0 to 0.2 for demethylation, and 0.8 to 1 for methylated CpG. However, using these stringent criteria resulted in no CpG fitting these rules, thus I decided to relax to β-value range to 0 to 0.6 for selecting CpG sites within a demethylated CpG region.

From a total of 485,00 CpG array data, 26 CpGs were selected using this first strategy. The list of CpGs and their details is presented in Appendix 8. The corresponding genes are listed below: (** indicates genes associated with 2 CpGs, **bold symbols** indicate the top 10 most disease related candidates)

- ***TNF*** (tumor necrosis factor), proinflammatory cytokine that has been implicated in a variety of diseases, including autoimmune diseases, insulin resistance, and cancer
- ***KSR1*** **(kinase suppressor of ras 1), related to RET signaling and RAS signaling pathway.
- *PBX2* (PBX homeobox 2), DNA-binding transcription factor activity and chromatin binding, pre-B-cell leukemia transcription factor
- ***TERT*** (telomerase reverse transcriptase), catalytic subunit of the enzyme telomerase
- ***IFITM1*** (interferon induced transmembrane protein 1),IFN-induced antiviral protein
- ***LOC100287036*** (Uncharacterized LOC100287036)
- *BCKDK* (branched chain keto acid dehydrogenase kinase), involved in BCKD the key regulatory enzyme of the valine, leucine and isoleucine catabolic pathways.
- ***TRAF5*** (TNF receptor associated factor 5), TNF receptor mediates TNF-induced activation
- *GPRIN3*** (GPRIN family member 3), G Protein-regulated inducer of neurite outgrowth
- ***PTPRCAP*** (protein tyrosine phosphatase receptor type C associated protein), associated with tyrosine phosphatase PTPRC/CD45, a key regulator of T- and B-lymphocyte activation
- *ANKRD11* (ankyrin repeat domain 11), modulates histone acetylation and gene expression in neural precursor cells, encoded protein inhibit ligand-dependent transactivation

- *GIMAP7* (GTPase, IMAP family member 7), the GTP-binding superfamily and to the immuno-associated nucleotide (IAN) subfamily of nucleotide-binding proteins
- *ITM2C* (integral membrane protein 2C), related to this gene include amyloid-beta binding, may play a role in TNF-induced cell death and neuronal differentiation
- *S1PR1* (sphingosine-1-phosphate receptor 1), involved signal of RAC1, SRC, PTK2/FAK1 and MAP kinases plays an important role in cell migration
- ***INS-IGF2*** (INS-IGF2 readthrough), related to AMP-activated Protein Kinase (AMPK) Signalling and Type II diabetes mellitus pathways
- *HPCAL1* (hippocalcin like 1),neuron-specific calcium-binding proteins family
- ***KRAS*** (KRAS proto-oncogene, GTPase), ras gene family, may plays a role in promoting oncogenic events by inducing transcriptional silencing of tumor suppressor genes
- *SEPTIN9* (septin 9), involved in cytokinesis and cell cycle control
- *BCL9L* (BCL9 like), Transcriptional regulator that acts as an activator promotes beta-catenin transcriptional activity and plays a role in tumorigenesis.
- *HLA-E* (major histocompatibility complex, class I, E), involved in immune self-nonself discrimination.
- *NCK2* (NCK adaptor protein 2), bind and recruit various proteins involved in the regulation of receptor protein tyrosine kinases.
- *PDE2A* (phosphodiesterase 2A), catalyse the hydrolysis of 3' cyclic phosphate bonds in the second messengers cAMP and cGMP
- *MICB* (MHC class I polypeptide-related sequence B), ligand for the NKG2D type II receptor activates the cytolytic response of natural killer (NK) cells, CD8 T cells.
- *ZBTB18* (zinc finger and BTB domain containing 18), transcriptional repressor of genes involved in neuronal development

This list was then manually narrowed down considering additional factors.

1. The methylation structure of the region surrounding the candidate CpG. For assay design, the primers need to bind with high specificity to a region of 20-30 base pairs around the candidate CpG. Therefore, the Candidate CpG site should not be an isolated CpG in order to provide sequence specificity for the methylation status. The array targets CpGs in CpG island however, those are spaced relatively widely for the probes to be able to bind efficiently. The methylation status of the neighbouring CpG should have a similar methylation status with the candidate CpG.
2. The effect size of the DM between HC and RA in naïve CD4+T-cell. Apart from the statistical significance, the actual difference in β value (Δβ) is also important. The bigger the Δβ between HC and RA, the better chance for the assay to pick up small difference by qPCR .
3. Finally, the biological relevance of the gene associated with the CpG candidate was also considered. Genes known to be related to the disease would undeniable have more credit on a ranking process.

From the 26 CpGs originally selected, the list was narrowed down to 10 candidate CpGs (highlighted in bold above) with the following ranking based on best fit to all rules from the top:

1. TNF	*2. IFITM1*	*3. TERT*
4. KSR1	*5. TRAF5*	*6. KRAS*
7. INS-IGF2	8. *PTPRCAP*	9. *LOC100287036*

5.3.1.2 Selection strategy 2: qMSP concept adding publicly available dataset

My second strategy was to extend the same principle to publicly available methylation datasets in T-cells and other population of cells, to enrich the analysis and notably to consider additional blood cells rather than only monocytes. 4 publicly available dataset were selected (detailed in Table 5-2) which contained methylation data in HC. All together data included

- naïve CD4+ T-cells,
- total CD4+ T-cells,
- NK cells, CD8 T-cells, B-cells, monocytes, as found in PBMC
- Neutrophil, Eosinophil, and Granulocytes included in WB

The Illumina 450K genome array analysis workflow was performed in a similar way for each dataset. The candidate CpG were selected using the same filtering rules.

1. Demethylation in the target cells: naïve, and total CD4+T-cells in HC sample. [β-value 0 to 0.6].
2. Methylation in the non-target cell; CD8+T-cells, NK cells, B-cells, monocytes, Neutrophil, Eosinophil, and Granulocytes in HC sample [β-value 0.8 to 1]
3. DM between HC and RA in CD4+T-cells [p-value ≤0.01, or Δβ > 0.10].

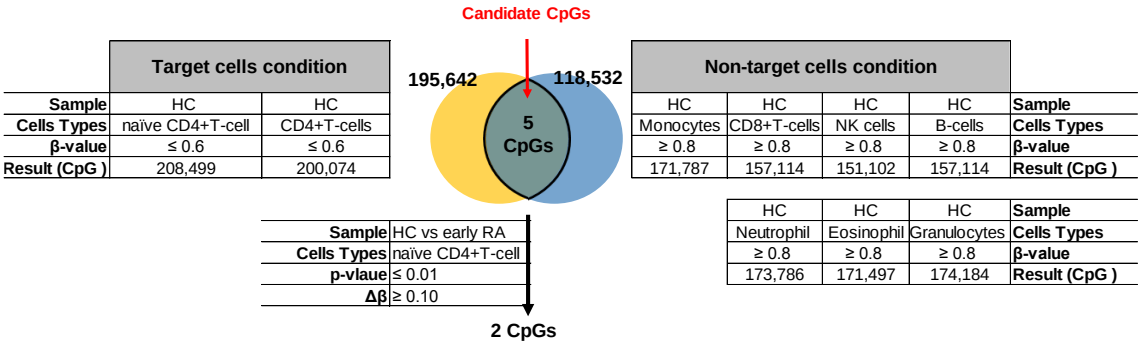


Figure 5-4 Results of the filtering criteria to select candidate CpG by my 2nd strategy

The filtering result workflow diagram is shown in Figure 5-4. CpGs associated with 5 genes were selected after applying the filtering criteria 1 and 2 (detail in the Appendix 8). Adding the 3rd filtering criteria considering whether DM between HC and early RA in naive CD4+T-cells, resulted in only 2 CpGs remaining on the list associated with the *RPTOR* and *ATP6V1H* genes.

- ***RPTOR*** (regulatory associated protein of MTOR complex 1), The encoded protein forms a stoichiometric complex with the mTOR kinase
- ***ATP6V1H*** (ATPase H+ transporting V1 subunit H),component of a multisubunit enzyme that mediates acidification of intracellular organelles necessary for protein sorting, zymogen activation, receptor-mediated endocytosis, and synaptic vesicle proton gradient generation.
- *AP5Z1* (adaptor related protein complex 5 subunit zeta 1), involved in homologous recombination DNA double-strand break repair (HR-DSBR).
- *RERE* (arginine-glutamic acid dipeptide repeats), co-localizes with a transcription factor in the nucleus, and its overexpression triggers apoptosis, may associates with histone deacetylase
- *CD40LG* (CD40 ligand), a member of TNF superfamily, notably expressed on activated CD4+ T-cells

5.3.1.3 Selection Strategy3: Using Delta Beta (Δβ) value in our dataset and publicly available dataset

In my 3rd Strategy, I selected candidate CpGs based on the size effect of the DM (i.e. different in methylation levels between HC and RA, the Δβ) in T-cells. Thus, it did not consider the methylation level (β-value) itself but focus on the difference (Δβ) between HC and RA.

I acquired publicly available datasets that contained methylation level data in HC and RA from several cell types (i.e. all PBMC) as well as form other form of autoimmune / inflammatory diseases (i.e. SLE, Systemic sclerosis (Ssc), Multiple sclerosis (MS) although unfortunately none were available for diseases that can easily be confused with RA in the an EAC. 8 datasets were obtained (detail described in Table 5-2). Of note, the RA patients in these datasets were from different stage of the disease as there is no other early RA dataset available. The information used was therefore from establish RA patient (236, 312, 380, 381). The Illumina dataset was analysed using a similar workflow as mention earlier for Δβ

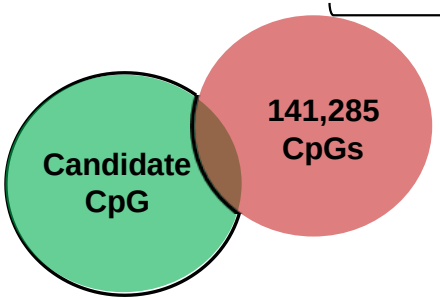
The candidate CpGs were selected by the filtering steps below:

1) Selecting CpG with $\Delta\beta \geq 0.10$ and p-value ≤ 0.01 between HC and RA in naïve CD4+T-cells
2) Keeping CpG with $\Delta\beta \geq 0.05$ and p-value ≤ 0.05 between HC and RA in (i) total CD4-T-cells and also (ii) in PBMC
3) filtering out CpG with $\Delta\beta \geq 0.10$ or p-value ≤ 0.01 between HC and RA in NK cells, B-cells, monocytes
4) Excluding CpG with $\Delta\beta \geq 0.10$ or p-value ≤ 0.01 between HC and IA (SLE, Ssc, and MS) in (i) PBMC or (ii) WB as non-RA specific

	Target DM condition			
Sample	HC vs early RA	HC vs RA	HC vs RA	HC vs RA
Cells Types	naïve CD4+T-cell	naïve CD4+T-cell	CD4+T-cells	PBMC
Δβ	≥ 0.10	≥ 0.05	≥ 0.05	≥ 0.05
p-vlaue	≤ 0.01	≤ 0.05	≤ 0.05	≤ 0.05
Result (CpG)	1,508	204	549	824

Non-target DM condition				
HC vs RA	HC vs early/est RA	HC vs SLE	HC vs Ssc	HC vs MS
B-cells	Monocytes	PBMC	PBMC	WB
≥ 0.10	≥ 0.10	≥ 0.10	≥ 0.10	≥ 0.10
≤0.01	≤0.01	≤0.01	≤0.01	≤0.01
9,963	19,829	66,592	100,204	4,122

141,285 CpGs



Target DM CpGs - Non target DM CpGs = Number of cadidate

	Combination of dataset	Number of Target DM CpG	Number of Non-Target DM CpG	Number of candidate CpG
Result 1:	All Target DM dataset	0	141,285	**0**
Result 2:	naïve CD4 T-cell (early RA) +CD4+T-cells + PBMC	1		**1**
Result 3:	naïve CD4 T-cell (early RA) + PBMC	27		**11**
Result 4:	naïve CD4 T-cell (early RA) +CD4+T-cells	27		**12**

Figure 5-5 Results of the filtering criteria to select candidate CpG by my 3rd strategy

Selection using these criteria and the datasets above resulted in no CpG fitting the rules (Figure 5-5, result 1) mainly because selected list of DM CpGs in naïve CD4+T-cell (from two datasets; early RA and established RA), total CD4+T-cell, and PBMC did not share any candidate. The filtering criteria needed to be relaxed.

The datasets with highest priority in this analysis is data from PBMC because it is the most desirable type of sample for such assay. It needs to be balanced against the 2nd important criteria which is to be selected from naïve CD4+T-cell from early RA patient as it is the only dataset from the type of patient that is relevant to the question asked (i.e. diagnostic). Indeed, to develop a diagnostic biomarker, it is important to detect the changes that occur in the early stage of the disease, while heterogeneity is likely to occur over the course of disease, further complicated by different treatment used and responded to or not. To illustrate this, the direct comparison of DM CpG from early (my dataset) versus established RA (American research group dataset (312), personal communication with access to raw data granted) in purified naïve CD4+ T-cells showed only 44 CpGs in common.

A series of different dataset combination for selecting candidate DM CpG was performed and the results are showed in Figure 5-5.

- Result 2: Selecting DM CpG common to naïve CD4+T-cell (early RA), CD4+T-cell, and PBMC based on [Δβ + p-value] for target DM CpGs gave 1 candidate CpG associated with
  - ***IRF8*** (interferon regulatory factor 8), transcriptional activator or repressor binds to the upstream regulatory region of type I IFN and IFN-inducible MHC class I genes

- Result 3: Selecting DM CpG common to naïve CD4+T-cell (early RA) and PBMC based on [Δβ + p-value] for target DM CpGs gave 11 candidate CpGs (full list described in Appendix 8.), the top 5 most relevant genes being;
  - ***HDAC4*** ***(histone deacetylase 4), alters chromosome structure and affects transcription factor access to DNA
  - ***MIR21***** (microRNA 21), involved in post-transcriptional regulation of gene expression (most validated target are tumour suppressors)
  - ***PSMB9*** (proteasome 20S subunit beta 9), An essential function of a modified proteasome, the immunoproteasome, is the processing of class I MHC peptides

- *PTMA* (prothymosin alpha), may mediate immune function by conferring resistance to certain opportunistic infections
- *S100P* (S100 calcium binding protein P), involved in the regulation of a number of cellular processes such as cell cycle progression and differentiation

Note; ** genes associated with 2 CpGs
*** genes associated with 3CpGs

- Result 4: Selecting DM CpG common to naïve CD4+T-cell (early RA) and CD4+T-cells on [Δβ + p-value] for target DM CpGs gave 12 candidates CpG (full list described in Appendix 8) the top 5 most relevant genes;
  - ***STAT5A*** (signal transducer and activator of transcription 5A), mediates cellular responses to the cytokines and other growth factors.
  - *PPTC7* (protein phosphatase targeting COQ7), T-Cell Activation Protein Phosphatase 2C
  - *ZBTB17* (zinc finger and BTB domain containing 17), ranscription factor involved cell cycle progression and plays a critical role in early lymphocyte development
  - *NAMPT*(nicotinamide phosphoribosyl transferase), involved NAD+ biosynthesis, reported to be a cytokine (PBEF) that promotes B cell maturation and inhibits neutrophil apoptosis.
  - *NCK2* (NCK adaptor protein 2), may involve in the regulation of receptor protein tyrosine kinases pathways and IL-2 Pathway

5.3.1.4 Concept discussion and decision making as to which candidate to pursue.

All strategies for the selection of candidate CpG shared a common focus on the methylation difference between HC and early RA in naïve CD4+T-cells which were showed to play major role in an early stage of disease development. It was therefore based on the hypothesis that such methylation change occurred early in RA and are related to the disease pathology.

Strategy 1 privileged my original dataset, being highly specific for naïve CD4+T-cells DM between HC and early RA, but with the major caveat that a qMSP assay may not work on mixed cell population as the DNA methylation status at the candidate CpG may be “diluted” in the template DNA without naïve CD4+T-cell purification or at least total CD4+ sorting before DNA extraction. Mitigating for this by filtering out CpG that were NOT highly methylated in other cells types was nonetheless part of the strategy, which following additional consideration for primer design and resulted in a list of 10 candidates, of which some clearly had been related to the disease before. This design should allow a qMSP to be successful on DNA extracted from CD4+T-cell and potentially from PBMC but less likely on WB.

Strategy 2 attempted to improve the detection of methylation signal from T-cells while working with a mix population of cells by using more data resources for other cell types to select CpG with a methylation status that was specific to T-cells compared to other cells in PBMC. This generated a list of 5 candidates, that did not seem very relevant to the RA disease pathogenesis (except maybe CD40L) but maybe over biased towards the lineage indeed (i.e. CD4+T-cells).

Strategy 3 worked differently from strategy 1 and 2. It simply selected candidate CpGs based on the difference in methylation levels (Δβ) between HC and RA in T-cells being large enough so that it could still be observed when diluted (i.e. in PBMC/WB). Thus, candidate CpG are those with large Δβ in naïve-CD4+T-cells, total CD4+T-cells, and PBMC while not showing Δβ in other cell types (i.e. CD8, B or NK cells) and not showing Δβ in other diseases. Ideally only early RA methylation data should be used and clearly established RA data had to be disregarded. Altogether, the strategy 3 gave a list of 22 candidate CpGs.

Obtaining sufficient methylation information to support the analysis according to these strategies was a major limitation. The results from strategy 1 used only my methylation dataset as resource for selection, but was highly selective For strategy 2, aimed for WB (or PBMC) as input material, which needed the methylation status of the candidate to be 100% methylated in all other types of cells to confirm that the demethylation in T-cells was really distinct between HC and RA. This was extremely limiting as the list of gene fitting these rules was maybe ("too") highly CD4+T-cell specific and no longer disease related.

The limitation for strategy 3 analysis was the lack of data truly relevant resources as ideally, it needed data of both HC and early RA in many cell types also PBMC or WB to account for the dilution effect in DNA from mixed cell population and also needed data of other early IA (rather than SLE and MS) to exclude DM shared with other IA, hence not absolutely specific for RA. There was too few datasets available to make this analysis as perfect as I would have wanted.

There was no overlap between the results of strategy 1 and 3 suggesting that they restricted the selection process in different ways bringing different strength to each selection list.

8 candidates CpGs from 3 strategies were chosen for the next steps, for target verification and assay development as listed below :

- ***TNF*** : The methylation pattern of the candidate CpG associated with the TNF gene (and the neighbour CpG +/- 500 bp from the array data) showed ideal methylation pattern for assay design. the region was demethylated in naïve CD4 T-cells (median β-value= 50%) while methylated in monocytes (86%). The DM between HC and RA in naïve CD4+T-cells was also high (Δβ= - 22%). This is also a promising target as TNF involvement with RA is well known. Furthermore, I had already confirmed the DM by bisulfite sequencing (Thesis result part1,) in total CD4+T-cells in early RA patient in the overall region surrounding the candidate CpG.
- ***IFITM1*** : Methylation pattern at candidate CpG of (median β-value = 52% for naïve CD4+T-cells and β= 86% for monocytes HC) and the nearby CpG were good for assay design. The DM between HC and RA in naïve CD4+T-cells is also high (Δβ= -20%).
- ***RPTOR*** and ***ATP6V1H*** : The methylation pattern of the 2 CpG associated with *RPTOR* and *ATP6V1H* meet all criteria of strategy 2and also show

good DM in naïve CD4+T-cells (Δβ = - 14% for RPTOR and - 21% for ATP6V1H.

- ***IRF8***: *IRF8* obtained from the most strictness criteria that DM in naïve CD4-T-cells (Δβ = 14%), CD4+T-cells (Δβ = 7%) and PBMC (Δβ = 5%).
- ***HDAC4***: *HDAC4* associated with 3 candidate CpGs. This CpGs located in a close proximity location in the genome and have high Δβ between HC and RA in naive CD4+T-cells (average Δβ = -15%) and PBMC (average Δβ = -7%) which could facilitate the biomarker assay design to detect the methylation change.
- ***MIR21***: associated with 2 candidate CpGs, DM in naïve CD4+T-cells (average Δβ = -16%) and PBMC (average Δβ = - 6%), and 1 CpGs DM in naïve and total CD4+T-cells but not PBMC.
- ***PSMB9***: show good DM in naive CD4+T-cells (average Δβ = -21%) and PBMC (average Δβ = -9%). The neighbour CpGs (+/- 500 bp from the candidate CpG) show similar methylation pattern.

The visual display of the methylation pattern in different cell types for selected candidates (and the CpG in +/- 1000 bp proximity) are showed as heatmaps of β-value and as Δβ (Figure 5-6 and Figure 5-7).

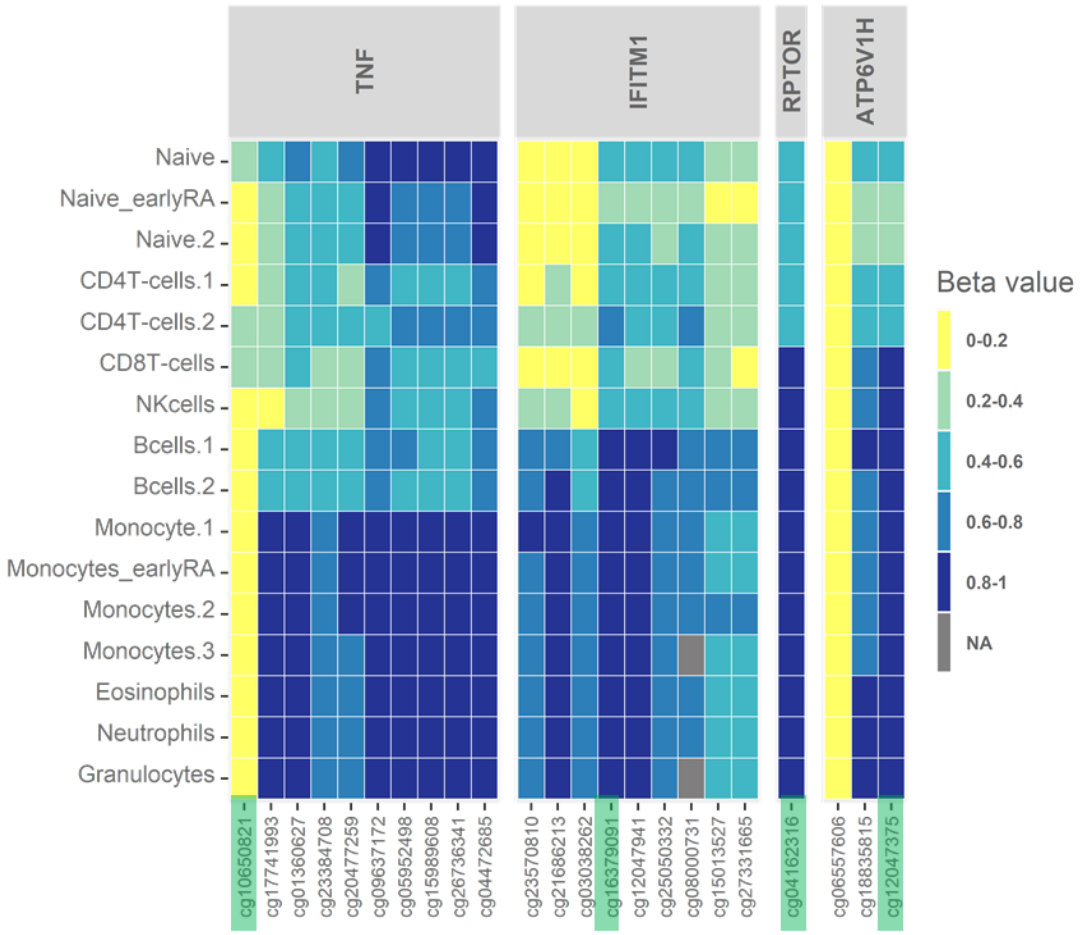


Figure 5-6 Heatmap of illumina dataset methylation level (β value) in candidate genes selected from strategy 1 and 2in different cell types for the top candidate CpG (highlight in green) and +/- ~1000 bp neighbour CpG. The methylation levels (beta-values) display from yellow for no methylation to blue high levels (as shown in the legend on the right side). The candidate CpGs/genes selected from strategy 1 (*TNF* and *IFITM1*) and 2 (*RPTOR*, *ATP6V1H*) should show low methylation level (0-0.6) in the target cells; naïve CD4+T-cells and CD4+T-cells and have high methylation level (0.8-1) in the non-target cells; monocytes and CD8+T-cells, NK cells, B-cells, and granulocytes.

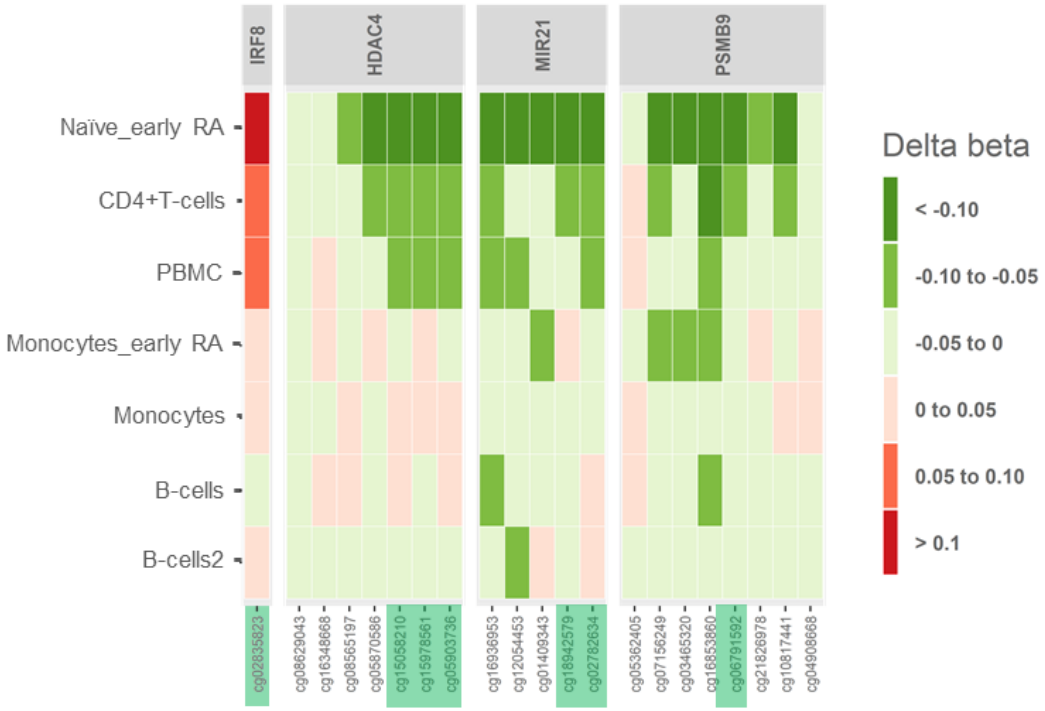


Figure 5-7 Heatmap of illumina dataset methylation differences (Δβ) in candidate genes selected from strategy 3 between HC and RA patients from different cell types of top candidate CpG (and +/- 1000 bp neighbour CpG). The Δβ display from dark green for the hyper-methylation (more than 10% difference in Δβ) to dark red for hypo-methylation (more than 10%) in RA. The candidate from strategy 3 (*IRF8, HDAC4, MIR21, and PSMB9*) are expected to have high Δβ (dark colours) in naïve CD4+T-cells and PBMC or CD4+T-cells, altogether with low Δβ (light colours) in other cells type.

5.3.2 Target verification : Bisulfite sequencing validation for the *IFITM1*

The methylation status of the region surrounding the candidate CpGs selected by the analysis should be verified to develop the assay. The illumina DNA methylation data of CpGs related to *IFITM1* was plotted against the position in genome (Figure 5-8). This allowed to define a region of 232 bp for sequencing .

The assay was optimised for this target region amplification and sequencing steps using fully methylated bisulfite converted control DNA. The successful assay was then used to access methylation in HC/patient samples in different cell types.

5.3.2.1 Bisulfite sequencing assay optimisation

Primer design

The Forward and reverse primers were designed to amplify the sense strand of bisulfite converted DNA from the CpG-island of the *IFITM1* gene in the region of ±150 bp from the candidate CpG. The expected PCR product was 232 bp covering 15 CpGs (Figure 5-9). By the rules for primer design for bisulfite sequencing, primers should amplify the region regardless of its methylation status so they should not contain any CpG. However, because of the location of the candidate CpGs in the island, it was difficult to design primer without CpG site. One unavoidable CpG was allowed in the reverse primer 3' end. The sequence and other details of primers and expected PCR product were described in the Appendix 5.

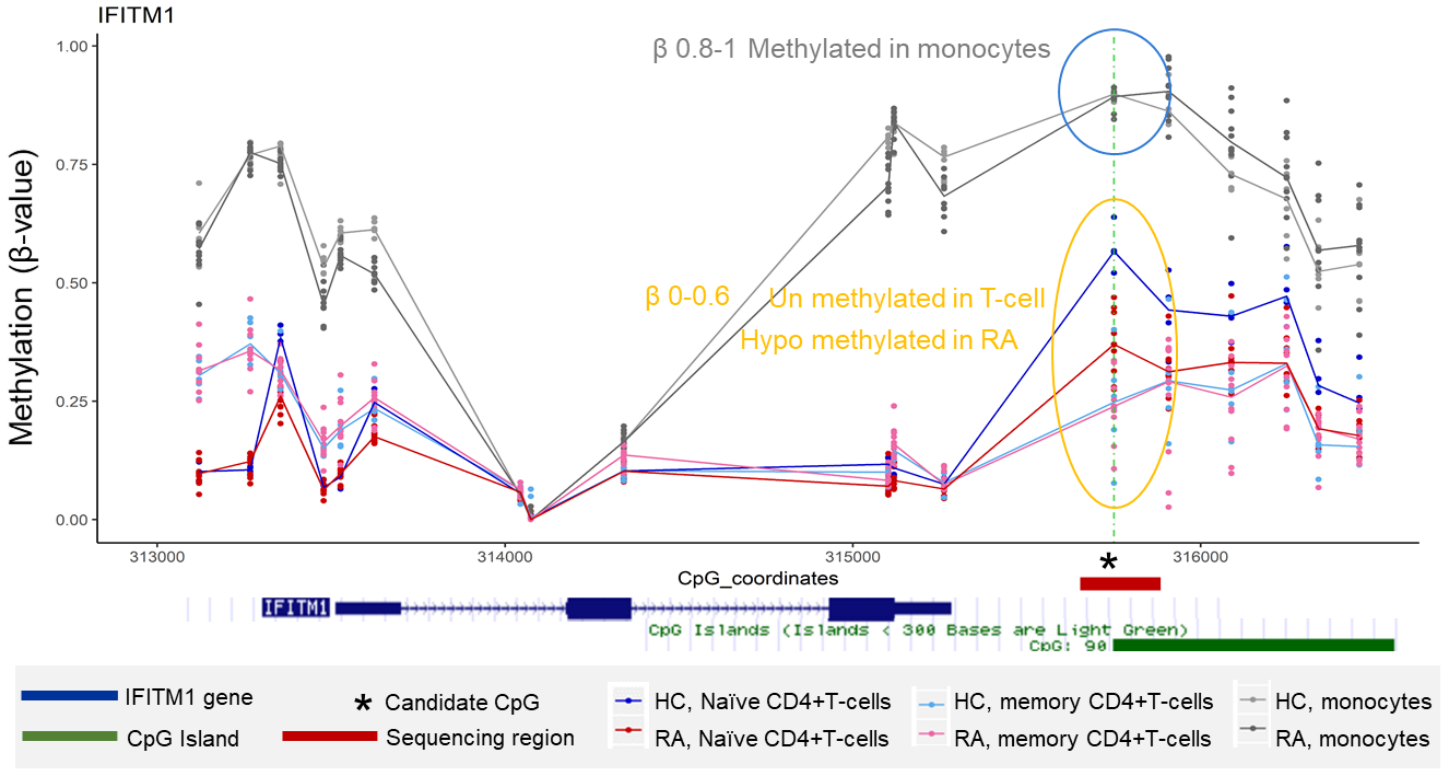


Figure 5-8 Illumina DNA methylation data (GSE121192) of the CpGs associated with *IFITM1* in different cell types. The methylation levels of an individual HC/RA patients (dots) against CpG coordinates on Chromosome 11 are plotted. The line graph shows the median of the methylation level of each sample group. At the candidate CpG (*), monocytes show fully methylated DNA while naïve T-cells show demethylation and DM between HC and RA groups.

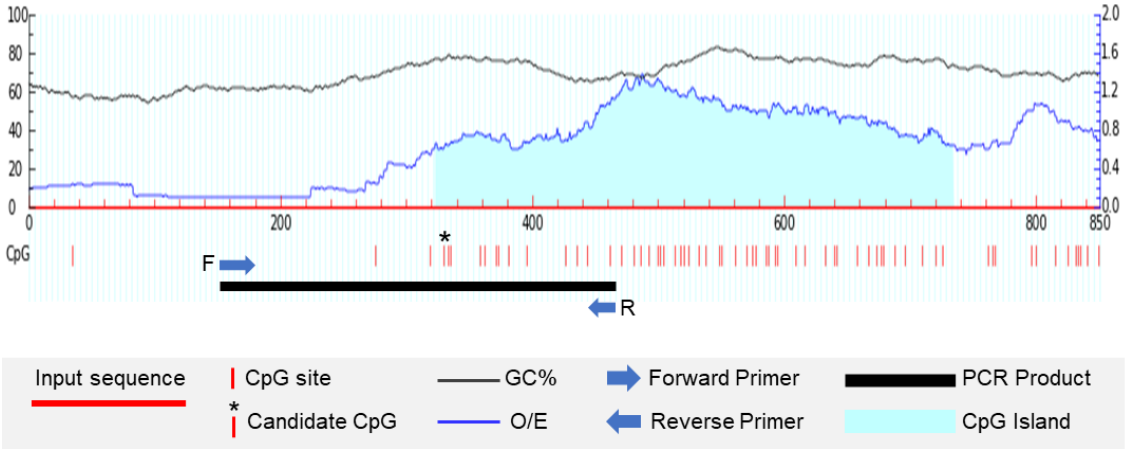


Figure 5-9 Bisulfite sequencing primers position on the *IFITM1* gene sequence. | represents an individual CpG position on the sequence. ***** marks the position of candidate CpG. Forward and Reverse primer present in blue arrow. The expected PCR product is 232 pb presented in bold black line.

Optimisation of the PCR amplification

Amplification of the target and the purity of the PCR product before the sequencing reaction are important. PCR reaction was adjusted from the standard PCR mixture and conditions to obtain a specific PCR product with a good quantity. Range of annealing temperature, Mg2+ concentration, and primer concentration were optimised to obtain the optimum PCR condition. The optimisation was perform using bisulfite converted methylated DNA control.

- Annealing temperature

Temperature ranging 58-61°C were tested (Figure 5-10,A). Agarose gel electrophoresis showed the correct product size (232 bp). Higher aneling temperatures increased the reaction efficiency up to 60°C and then reduced it.

- Mg2+ concentration

Increasing Mg2+ concentrations between 1.5, 2, 2.5, and 3 mM were then tested (Figure 5-10,B). A 2 mM Mg2+, is the lowest concentration that produces a high amount of PCR product and was chosen to use in PCR reaction.

- Primer concentration

Increasing primer concentrations between 0.125, 0.25, 0.5, 0.75, and 1 mM were then tested (Figure 5-10,C). Primer concentration at 0.25 mM produces the highest DNA band intensity and was chosen to use in PCR reaction.

The final fully optimised conditions for the amplification of the *IFITM1* target CpG island region are described in the Table 5-3.

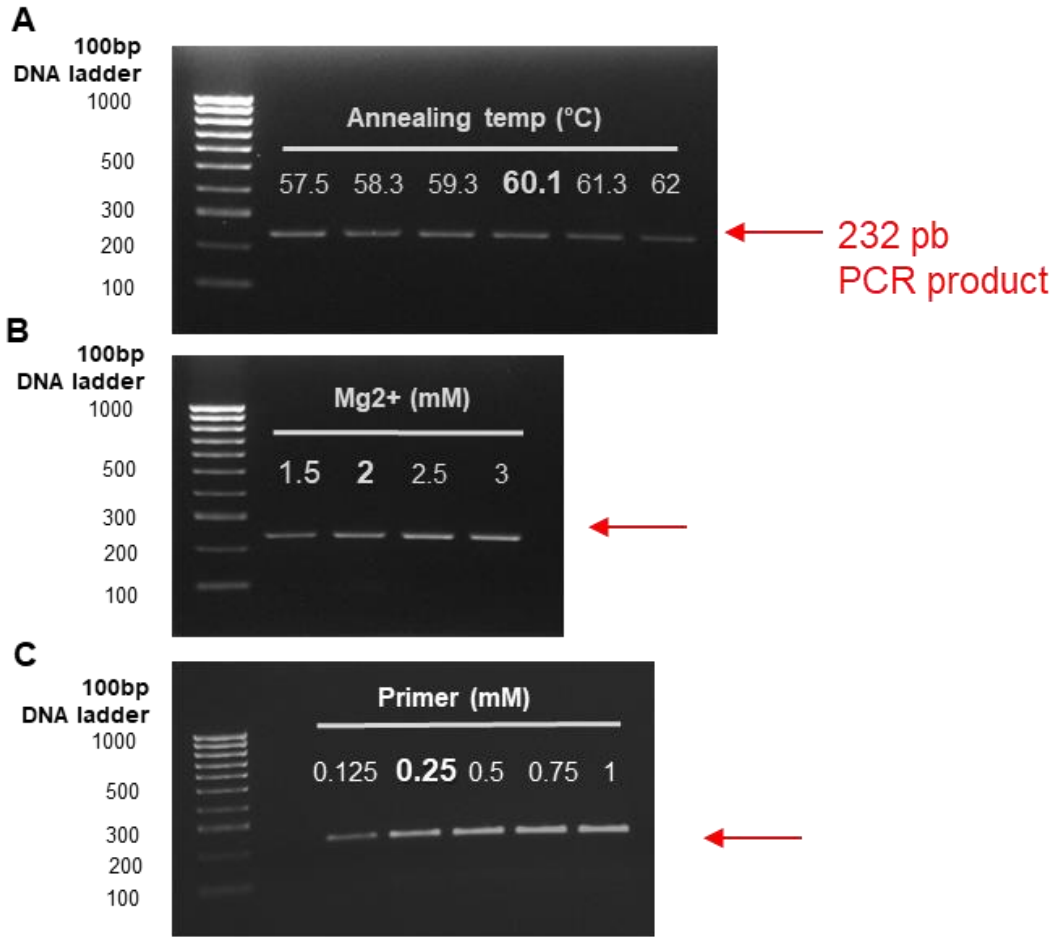


Figure 5-10 Agarose gel electrophoresis showing PCR product of the *IFITM1* gene amplicon for sequencing A) at various angling temperature and B) Mg2+ concentration , and C) primer concentration. The temperature, Mg2+ and primer concentrations that are bold indicated as the selected condition.

Table 5-3 PCR Conditions and cycles for the *IFITM1* gene amplification.

Stock conc.	Reagent	Final conc.	vol/1 reaction (uL)
10X	PCR Buffer*	1x	2
25mM	MgCl2	2 mM	0.4
10 mM of each	dNTP mix	200 uM of each	0.4
2.5 uM	Forward Primer	0.25 uM	2
2.5 uM	Revers Primer	0.25 uM	2
5 Unit/ul	HotStartTaq DNA polymerase	2.5 U/reaction	0.5
	Distilled water		10.7
	Template DNA	< 1ug/100 ul reaction	2
	Total Volumn		20

*Buffer contain 1.5 mM

PCR Cycle			
Initial denaturation	95ºC	15 min	
Denaturation	94ºC	10 s	40 cycles
Annealing	60ºC	20 s	
Extension	72ºC	45 s	
Final extension	72ºC	10 min	

Sequencing reaction optimisation

After successfully amplifying the *IFITM1* target region and cleaning the product, the sequencing reaction was performed. The pre-optimized sequencing condition previously used for the TNF gene failed to produce a readable sequence for the *IFITM1* target product (Figure 5-11,A). Further optimisation was needed for this target.

Various factors may affect the efficiency of reaction including (but not limited to) the enzyme, primer or template concentrations. Varying conditions for these factors was tested. The best condition was chosen based on the raw intensity, the shape of raw intensity, or/and the electropherogram peak. The control sequencing reaction (using the company template DNA and primer) was also performed as a reference for a good sequencing example (Appendix 9).

- Primer concentration

Various primer concentrations at 40, 80, 160 nM were tested. Considering the raw intensity, reducing primer concentration slightly increased the sequencing signal intensity. However, only the beginning of the PCR product could be sequenced. (Appendix 9).

- DNA Template concentration

Various amounts of PCR product (0.5, 1, 2, 4 µL) were tested. Lower amount of DNA template slightly increased the signal intensity. However, only the beginning of the PCR product could be sequenced again (Appendix 9). Adjusting the DNA template and primer concentration at this state did not show major improvement.

- Sequencing enzyme

Various enzyme concentrations between 0.0625, 0.125, and 0.25X of the ready reaction mix were tested. Increasing enzyme concentration notably helped improve the signal. At 0.25X concentration, the whole length of the PCR product could be sequenced. Using this concentration with the primer and template optimal concentration, the electropherogram with well-defined peaks and good signal-to-noise ratios were obtained (Appendix 9).

The final sequencing conditions are summarised in Table 5-4. Example of good quality sequencing results following full optimisation are displayed in Figure 5-11, B.

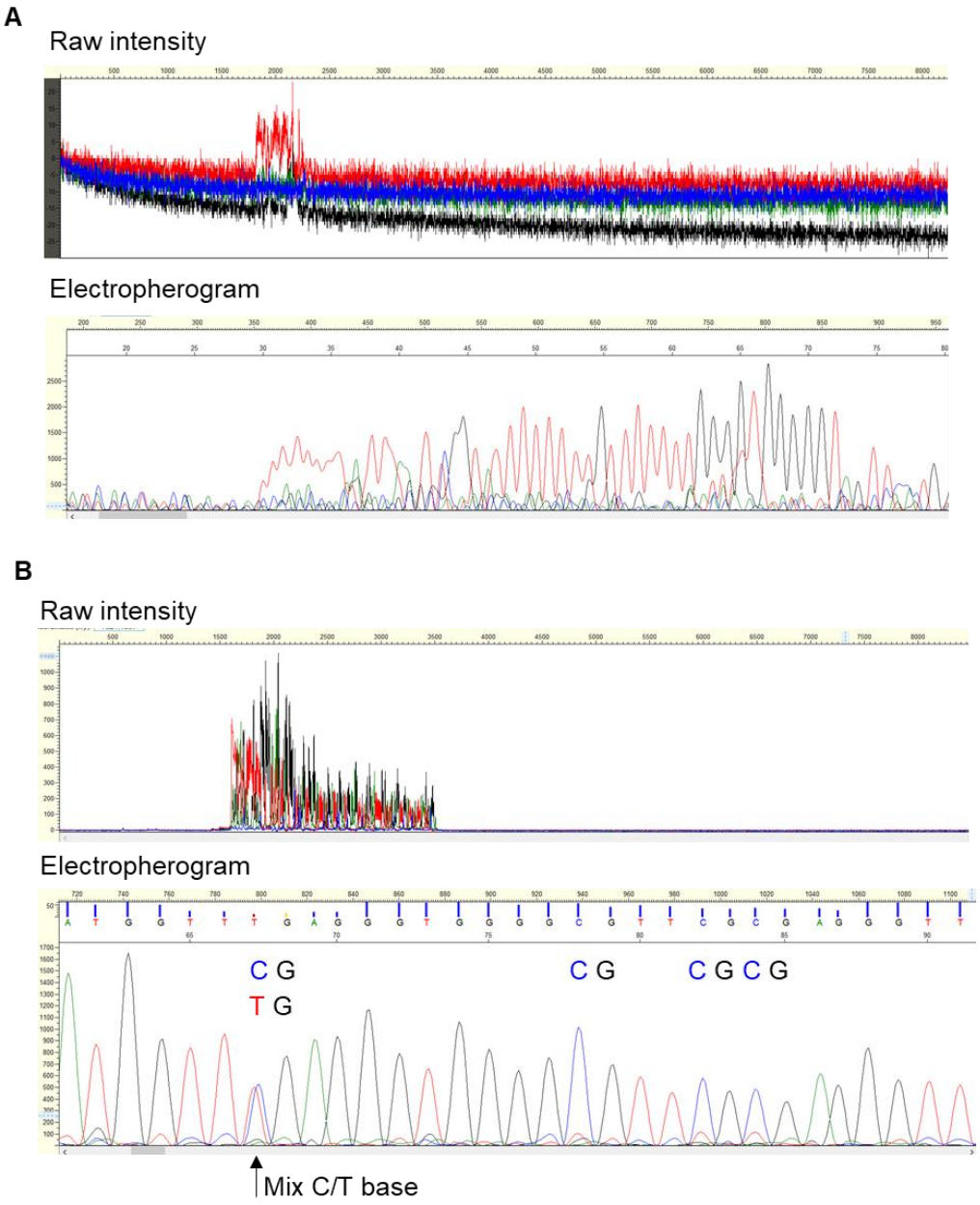


Figure 5-11 Sequencing raw intensity and electropherogram using the A) standard protocol of the *IFITM1* CpG island with poor sequencing data while in B) example of good quality sequencing results following full optimisation. Raw intensity shows the signal of overall sequencing product and no sign for signal saturation or low signal, dye blobs, and primer dimers. The electropherogram peak represents a single nucleotide in the DNA sequence. Each nucleotide showed in a different colour; A-green peaks, T- red peaks, C-blue peaks and G -black peaks. The good quality sequencing shows well-formed, distinctive single. coloured peaks. CpG positions were labelled. Methylated cytosine shows as CG and the unmethylated cytosine shows as TG. Arrow points an example of equally methylated and unmethylated cytosine.

Table 5-4 An optimised sequencing condition for *IFITM1* target region.

Stock conc.	Reagent	Final conc.	vol/1 reaction (uL)
2.5X	Ready Reaction Mix (RRM)	0.25X	1 ul
5X	ABI 5Xsequencing buffer	0.75 X	1.5 ul
0.4 uM	Primer (forward only)	0.04 uM	1 ul
	Nuclease free water		5.5 ul
	DNA Template		1 ul
		Total volume	10 ul

PCR Cycle			
Initial denaturation	96ºC	1min	
Denaturation	96ºC	10 sec	28 cycles
Annealing	50ºC	5 sec	
Extension	60ºC	4 min	
	Hold at 15ºC		

5.3.2.2 *IFITM1* bisulfie sequencing performed on patients sample

The candidate CpG in the *IFITM1* genes was selected based on a bioinformatic strategy. Ideally, the candidate CpG need to be demethylation in T-cell while fully methylated in other cells and it also need to be DM between HC and RA in T-cell. The candidate CpG methylation status needed to be checked before further use for a biomarker assay development. Bisulfite sequencing was used to access methylation status of different cells.

5.3.2.2.1 DNA methylation of 5 cell subsets

To ensure that the methylation status is demethylated in T-cells but not in other cells at our candidate CpG, the DNA of 5 cells subset was tested. CD4+T-cells, CD8+ T-cells, NK-cells, B-cells and monocytes were used for *IFITM1* bisulfite sequencing. PBMC from 3 healthy donors were collected and sorted into 5 cell subsets using FACS following cell surface staining for CD4, CD8, CD19, CD56, and CD14. Sorted cells purity was > 95%. DNA from all samples was extracted and bisulfite converted. The bisulfite sequencing of *IFITM1* gene was performed using the optimised conditions mention above. Methylation of the individual 14 CpGs out of 15 CpGs in the target region (232 bp) in individual cell subset was quantified as % of methylation and presented in Figure 5-12,A. The methylation of one CpG was not quantified due to the its position near the beginning of the sequence.

At the candidate CpG (2nd CpG), levels of methylation observed in monocytes (mean 96.53± SD 0.04%) were as expected and were similar to that of B-cells (97.28±0.63%). The sequencing result of methylation in CD4+T-cells (81.28 ± 5.65%), CD8+T-cells (84.48±4.31%) and NK cells (82.81±3.23%) were in the same range. These methylation level were higher than expected. This pattern appeared the same throughout the other 13 CpGs in the sequencing region. This results were not totally unexpected as when I retrieved data from publicly available sources, it showed that CD4 as well as CD8 and NK cells showed <60% methylation, (Figure 5-12,B). The methylation levels of theCD4+T-cells in this region were not sufficiently distinct from that of these 3 other cell types so that a signal coming from these cells could prevent the detection of methylation differences form the CD4T-cells. Thus this region of *IFITM1* is unlikely to be a good target for qMSP design.

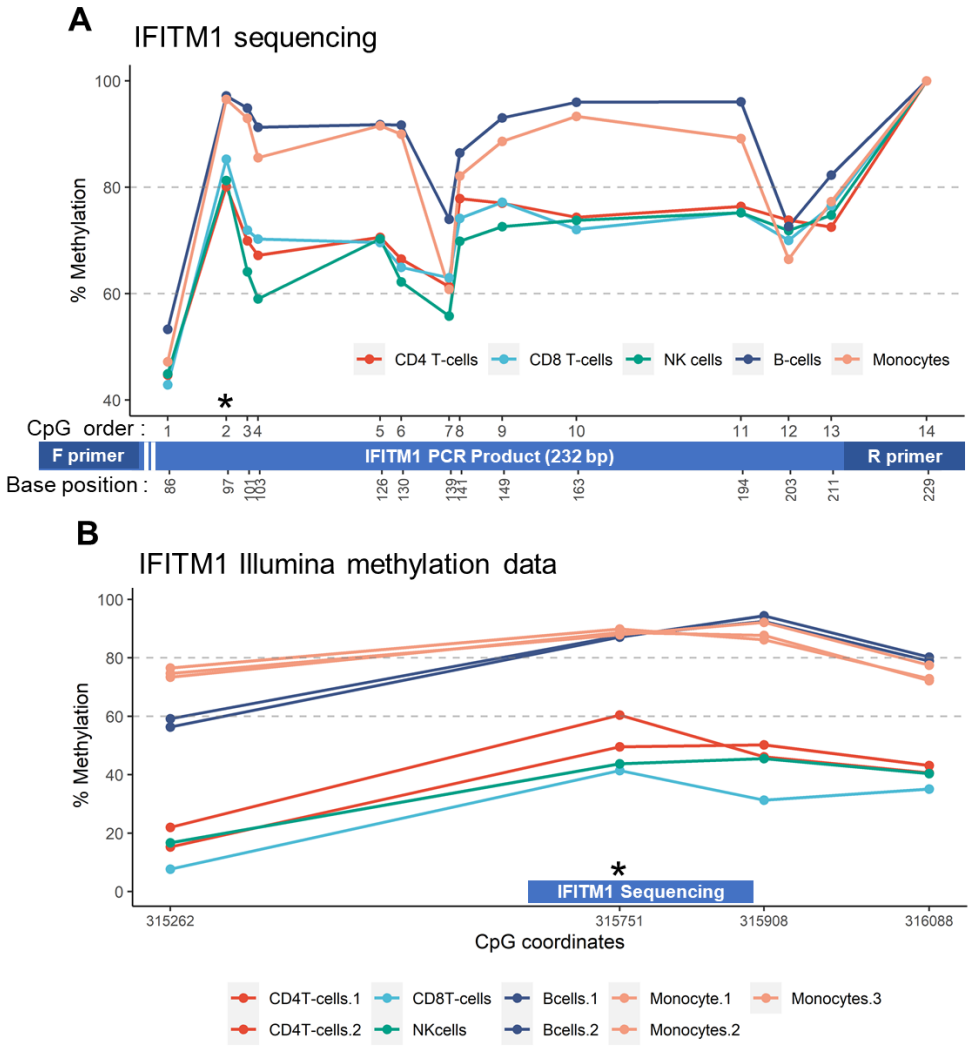


Figure 5-12 *IFITM1* DNA methylation

(A) Bisulfite sequencing result from 5 sorted cells subset of HC (n=3). The average percent of methylation at the individual 14 CpG in IFITM1 PCR product (232 bp) shows in the plot. The 2rd CpG which is the candidate CpG and the only CpG with illumine methylation data available is labelled with *.

(B) Publicly available Illumina methylation data of 5 cells subsets. The methylation levels the candidate CpG (*) and the approximal CpGs against CpG coordinates on Chromosome 11 are plotted. *IFITM1* sequencing region (blue box) is aligned to the chromosome (CpG coordinates 315,655 to 315,886). The sequencing region contain only one illumina CpG.

5.3.2.2.2 DM of the *IFITM1* gene between HC and RA in CD4+T-cells

Despite the limitation described above, the methylation analysis between HC and RA patients was performed using bisulfite sequencing from DNA of purified CD4+ T-cells. CD4+T-cell were isolated from 18 HC and 16 RA frozen PBMC by magnetic bead negative isolation. DNA was extracted, quantified and bisulfite converted and used in *IFITM1* bisulfite sequencing. Samples were part of the IACON cohort.

At the candidate CpG position (2nd CpG), methylation in HC and RA was 83.81 ± 6.41% and 80.40 ± 4.35%, respectively (Figure 5-13,A). The results therefore confirm a trend for hypomethylation in RA at the candidate position as well as and overall in all others CpGs of the region (except for last CpG). This is only statistically significant in 2 CpGs (5^{th} CpG p=0.045 and 9^{th} CpG p= 0.037) despite very small $\Delta\beta$-value (~ 4%) compared to what was observed in naïve CD4+T-cells.

An unexpected observation was made. That the results of the sequencing data at the candidate CpG site in HC showed a correlation between methylation and age, older age subjects tending to have lower methylation. (Figure 5-13, B, rho=-0.425, p-value=0.1159). This may reflect the overall accumulation of experienced memory cells with age in total CD4+T-cells.

RA patients had an age range different from HC (mean 40.08 ± SD14.55 year old for HC, 59.88 ±13.00 year old for RA). The lower methylation observed in RA patients may therefore be biased due to the effect age on methylation. Reducing the number of HC to the 11 oldest and keeping the youngest RA patients who were far outside the age range, a 2^{nd} analysis was performed(Figure 5-13, F). Over a similar age range, the overall levels of methylation of the candidate CpG no longer showed association with age. There was, however, no significant DM between HC and RA in the target region in CD4+T-cells although on average the methylation level (β value) was lower in RA. This bisulfite sequencing results suggest that additional confounding parameters would need to be considered when analysing data from real cohorts.

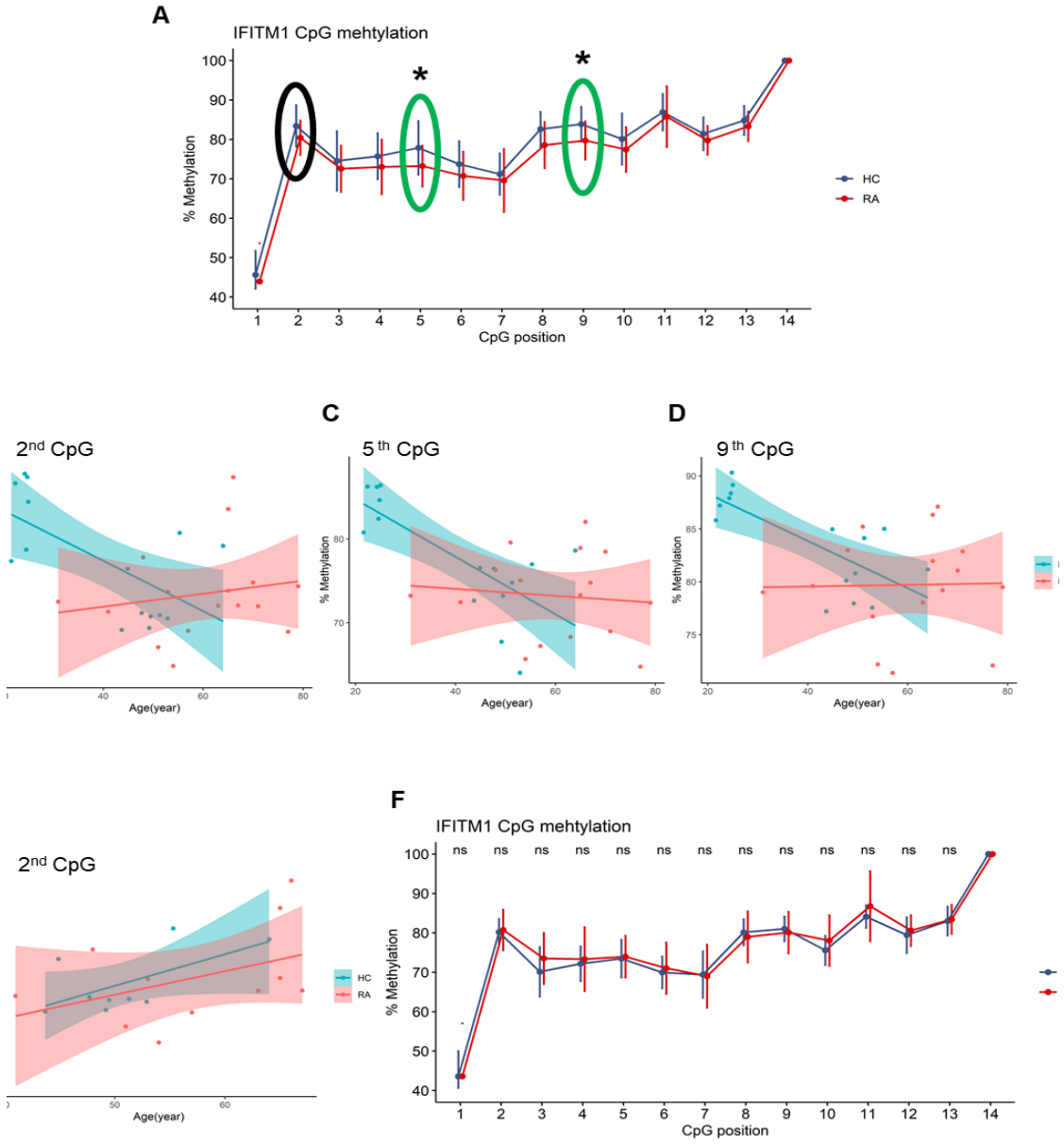


Figure 5-13 *IFIMT1* bisultfite sequencing result of CD4+T-cells.

A) The percent of methylation (mean ± SD) of HC and RA at the individual 14 CpGs on *IFITM1* region. The 2nd CpG is the candidate CpG (Black circle). Statistical analyses comparison of HC and RA were performed using the MWU test (*p<0.05, green circle).

B) The association of age and methylation at the 2nd CpG of HC and RA using the linear regression model of overall patient.

C) The association of age and methylation at the 5th CpG of HC and RA using the linear regression model of overall patient.

D) The association of age and methylation at the 9th CpG of HC and RA using the linear regression model of overall patient.
E) The association of age and methylation at the 5th CpG of HC and RA using the linear regression model of patient in age range 35 to 70.
F) The percent of methylation (mean ± SD) of HC and RA age range from 35 to 70 at the individual 14 CpGs on *IFITM1* region.

Altogether, *IFITM1* sequencing in 5 cell subsets highlight the issue in the design strategy if not using purified CD4+T-cells. Levels of DNA methylation in total CD4+T-cells were higher than observed in naïve CD4+T-cells creating a second issue. This confirmed that the principles under strategy 1 (i.e. to detect DM in T-cells while working with mix population of cells in PBMC or WB) was not optimal for *IFIMT1*. Strategy 2 which was design to mitigate such issue indeed did not retain *IFITM1*, confirming that the concepts associated with both strategies were sound . In addition, *IFITM1* sequencing results in HC and RA in CD4+T-cells showed no DM when results were adjusted for age. Based on these result, I decided that the *IFITM1* gene was not suitable to use for a full biomarker assay development.

At this point in my project and due to the difficulty and time consuming aspects of optimising multiple bisulfite sequencing assays and the limitation of my study time, I decided to give priority to candidate CpGs from strategy 2 and 3 moving forward to developing methylation detection by qMSP assay directly.

5.3.3 Development of qMSP assay for target genes

This section describes the development of an assay for detecting DNA methylation of candidate CpGs. Development of qMSP based on a SYBR green detection method was first attempted, however, the successful final assay was based on a TagMan detection method. Some of the candidates were not tested due to the nature of sequence surrounding the candidate CpG not allowing primer design after bisulfite conversion.

The *TNF* gene (top candidate from startegy-1) had confirmed DM by sequencing (284). Although there is evidence that the *TNF* methylation status in other cell subsets is going to interfere with the assay in mix cell population, the DM between HC and RA was quite significant over a relatively large region. Thus, I believe this gene was worth keeping on the list of assay development.

The design of assays was attempted for :

Strategy-1	Strategy-2	Strategy-3
TNF	*RPTOR*	*HDAC4, MIR21*

5.3.3.1 Optimisation of qMSP for unmethylated DNA target and internal control genes, using a SYBR-green based assay

Primer design

For the target gene assays, forward and reverse primers were designed to detect un-methylated CpG in the candidate genes. The internal control assay aimed to normalise the amount of input DNA was designed to avoid CpG site ensuring DNA methylation independent amplification. Two internal control assays were designed using *GAPDH* and *ACTB* genes. Designing good qMSP primers to meet all primer design rules is difficult. For some genes, ideal condition could not be met, and several version of the primer-set had to be designed (12 pairs). The sequences and other details of all primers were listed in the Appendix 5.

Optimisation of PCR condition

The assay optimization of target genes and internal control genes followed the same workflow. First, I began with primers optimisation to find the right concentration of primers giving the highest fluorescent intensity without promoting the formation of primer dimer or non-specific product. Then the primer specificity for un-methylated DNA (target gene), and the methylation independent amplification (control assay) were checked by comparing the amplification of reaction using 100% un-methylated and 100% methylated control DNA as input template. Amplification plot (Ct value), melting curve analysis, and agarose gel electrophoresis were considered together to determine the best PCR conditions. The final step was testing the reaction efficiency using dilution curves. Of 12 primer sets, some were dropped in the middle of the workflow. To avoid an overload of data, I decided to present in this part, the optimization of 2 primer sets as examples of representative optimisation,1 for target gene assay and the other for internal control assay.

Primer validation of target gene assay : *HDAC4* primer-set

3 primer concentrations (varied from 50-900 nM) of forward and reverse primers with 9 combinations were tested. Amplification plot and Ct value of each primer condition are show in Figure 5-14,A. The Highest primer concentration at F900/R900 nM gave the best amplification(Ct = 25.72). Melting curve analysis was performed after the PCR reaction to check the specificity of the amplification. A single peak (single temperature) in the melting curve analysis, which normally means one size of PCR product was observed, suggesting the absence of primer-dimer or nonspecific amplification, at all primer concentrations (Figure 5-14,B). Agarose gel electrophoresis was performed to visually inspect the PCR product and to confirm the amplification specificity. The results from agarose gel electrophoresis, however, contradicted the melting curve data (Figure 5-14,C). At higher reverse primer concentrations, primer-dimer/nonspecific products were observed in the gel while a single peak was present by the melting curve analysis. The fluorescent signal that was detected and used to obtain a Ct value was, in fact, the mixed-signal from the target DNA product and non-target PCR product. A negative control (No template control, NTC) was also performed at each primer condition. The agarose gel electrophoresis showed a double stranded DNA band at ~50 bp, suggesting primer dimers (as well as amplification of un-methylated DNA at ~200 bp, Figure 5-14,D). Although in an ideal reaction, non-

specific/primer-dimer should be absent in the no-template PCR reaction, in practice sometimes this might appeared as a result of consuming of PCR reagent resource on a short double stranded DNA product or resulting from annealing of primers on each other's.

Therefore, F50/R50 primer concentration that gave the right PCR product with more Ct difference between un-methylated DNA and no template control than the other conditions was kept forward. The melting curve analysis applied after SYBR green-based detection qPCR, for this primer pair, however, doesn't seem to be useful. A single peak for both the target PCR product and the non-specific product/primer dimer appeared at the same temperature in the melting curve plot, while showing a different size in agarose gel. This emphasize the need for using agarose gel electrophoresis.

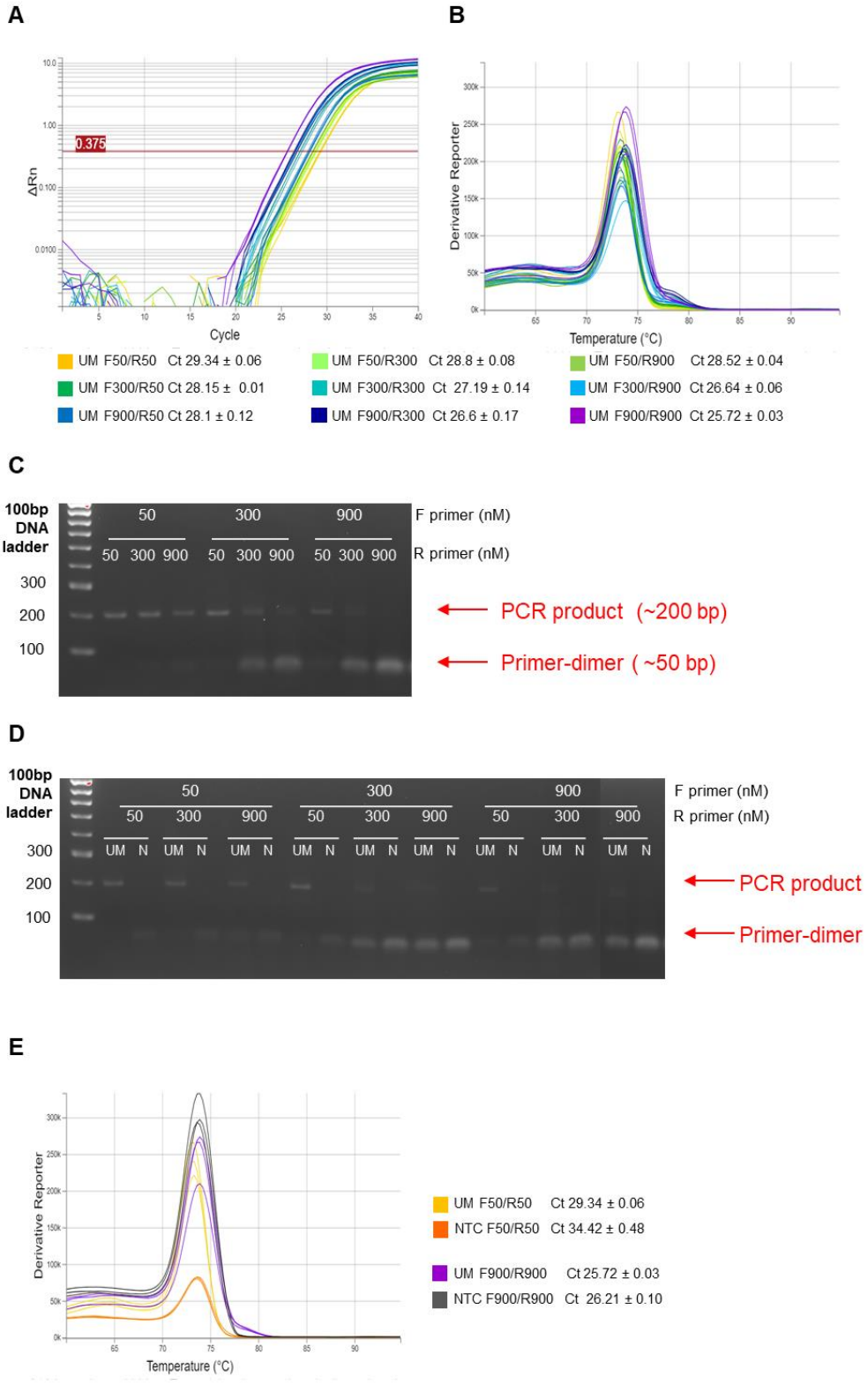

A
B
0.375
ΔRn
Cycle
Derivative Reporter
Temperature (°C)
UM F50/R50 Ct 29.34 ± 0.06
UM F50/R300 Ct 28.8 ± 0.08
UM F50/R900 Ct 28.52 ± 0.04
UM F300/R50 Ct 28.15 ± 0.01
UM F300/R300 Ct 27.19 ± 0.14
UM F300/R900 Ct 26.64 ± 0.06
UM F900/R50 Ct 28.1 ± 0.12
UM F900/R300 Ct 26.6 ± 0.17
UM F900/R900 Ct 25.72 ± 0.03
C
100bp DNA ladder
300
200
100
50 300 900 50 300 900 50 300 900
F primer (nM)
R primer (nM)
PCR product (~200 bp)
Primer-dimer (~50 bp)
D
100bp DNA ladder
50 300 900
UM N
F primer (nM)
R primer (nM)
PCR product
Primer-dimer
E
Derivative Reporter
Temperature (°C)
UM F50/R50 Ct 29.34 ± 0.06
NTC F50/R50 Ct 34.42 ± 0.48
UM F900/R900 Ct 25.72 ± 0.03
NTC F900/R900 Ct 26.21 ± 0.10

Figure 5-14 Primer optimisation of *HDAC4* gene. A) Amplification curve B) Melting curve with a single peak at 74^{o}C and C) agarose gel electrophoresis with 1 PCR product at ~200 bp size and primer dimer at <50 bp) of different concentration of forward and reverse primer using UM DNA control as template. D) different concentration of forward and reverse primer using UM DNA control and no DNA input (label in N). E) Melting curve for NTC showing same pick ~74^{o}C, as for the un-methylated DNA PCR product

Primer optimisation for Control gene assay: *ACTB*

The same strategy was applied, although additional primer concentrations were tested at F50/R100 and F50/R200 mM. The best concentration combination for amplification was F50/R200 mM producing the right PCR product (~200 bp) and no primer dimer in both 100% un-methylated DNA and NTC. Primer dimer were seen when using higher concentration primers on both the agarose gel (~50 bp) and as a 2nd peak on melting curve (arrow). The result from the amplification plot, melting curve analysis and agarose gel electrophoresis therefore agreed with each other for this primer pair (Figure 5-15).

As mention earlier, the same optimisation workflow was applied for other primer sets. In some case, optimising the annealing temperature was attempted to improve the result. Lowering temperature was aimed at facilitated amplification, and improved the Ct value and band intensity, while the higher temperatures were expected to help reduce the non-specific products. Unfortunately, in most cases varying the annealing temperature did not ameliorate results.

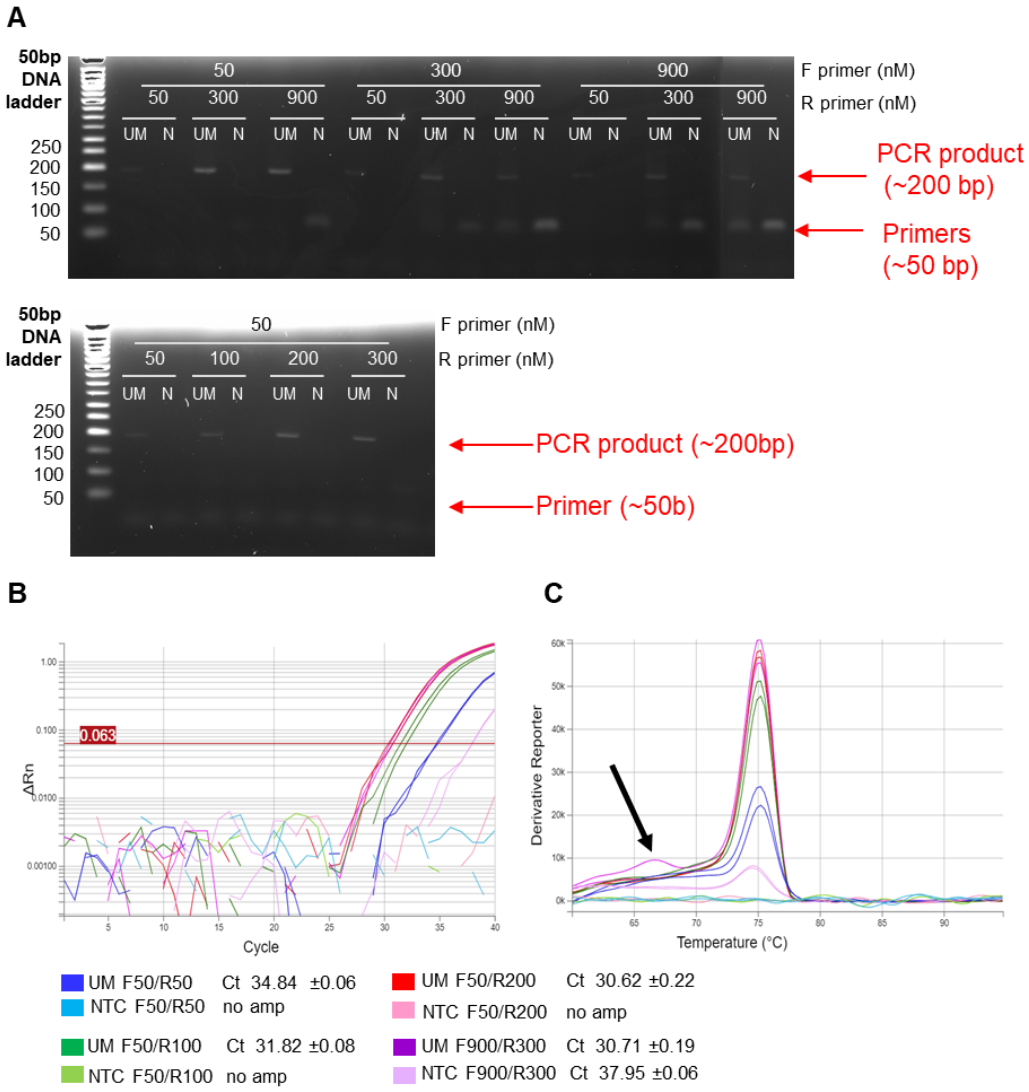


Figure 5-15 Primer optimisation of *ACTB* gene. A) agarose gel electrophoresis with 1 PCR product at ~200 bp size and primer dimer at <50 bp), B) Amplification curve, and C) Melting curve with a single peak or double peak (arrow pointed the 2rd peak) of different concentration of forward and reverse primer using UM DNA control and no DNA input (label in N) as template.

5.3.3.1.1 Specificity for un-methylated DNA

After obtaining the best working primer conditions for each primer set, the reaction was also tested using 100% methylated control DNA versus 100% un-methylated control DNA to ensure specificity to un-methylated target genes and ensure equal amplification (independent of methylation) for the internal control assay. The 12 primer pairs were tested. The results are presented on agarose gel electrophoresis picture (Figure 5-16).

For the target gene assays, the only primer set that was able to discriminate between un-methylated and methylated control DNA was *HDAC4*. For all other candidate genes, the designed primer sets were not sufficiently specific to un-methylated sequence and showed some levels of amplification for methylated DNA.

For internal control gene assays, *GAPGH* V1 and V2 primer sets could amplify methylated and unmethylated control DNA equally well, however, non-specific band appeared in the NTC. *ACTB* v1 was also able to equally amplify methylated and unmethylated DNA in the absence of primer-dimer (in both reactions with template and NTC).

At this stage, it appear that only *ACTB* and *HDAC4* could fulfil the condition for an optimal qMSP assay using SYBR green.

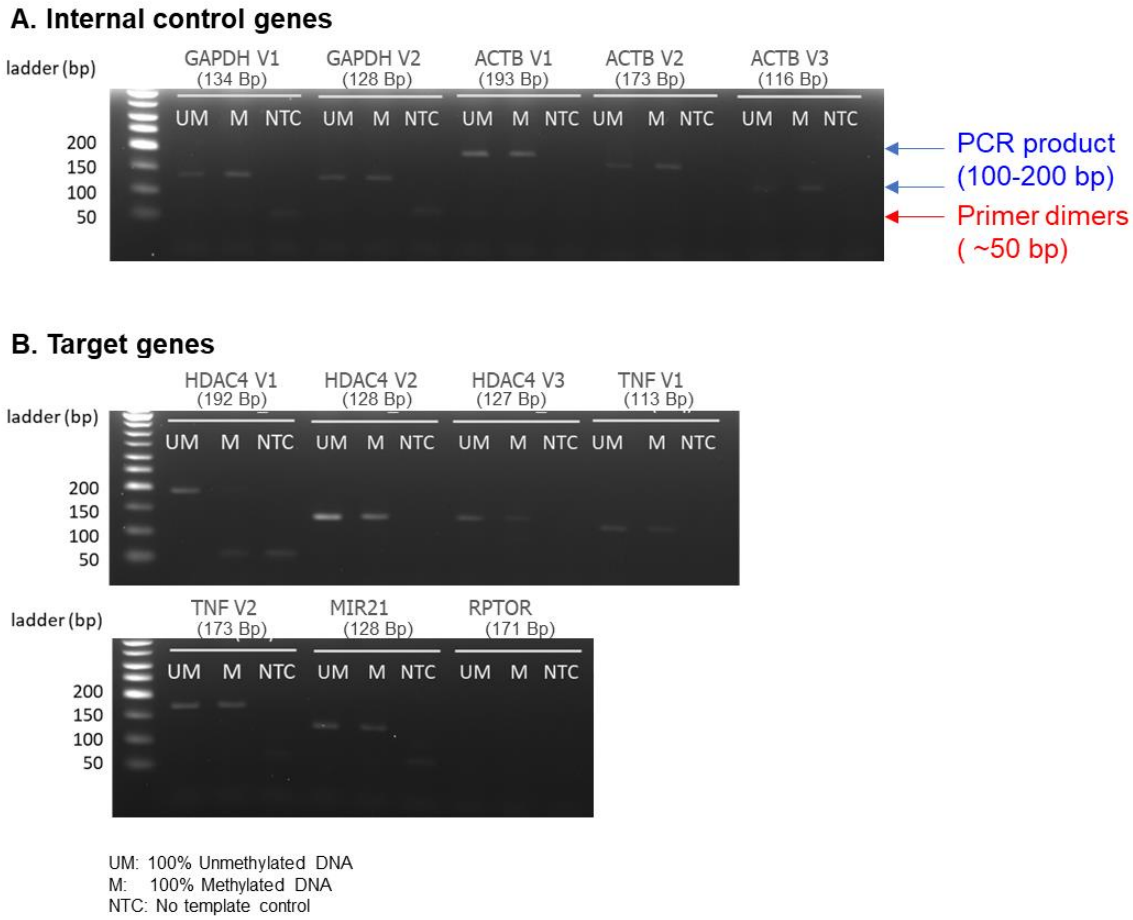


Figure 5-16 Agarose gel electrophoresis shows assay specificity of A) Internal control genes and B) Target genes of each primer set using different DNA templates; UM (100% un-methylated control DNA), M (100% methylated control DNA), NTC (no template control).

5.3.3.1.2 SYBR green qPCR efficiency

The last step for assay optimisation was to test qPCR reaction efficiency. Standard curve generated by serial dilution of DNA template was performed for 2 primer genes (*ACTB* and *HDAC4*). For the individual assay, dilution curve was plotted against template concentration and it's Ct value and was then fitted to a linear regression model. Assay efficiency was calculated from the slope of the regression line (Figure 5-17).

The efficiency of all assay was over 150% (i.e. more than 2 PCR product by cycle) which is not an acceptable efficiency. Too high/low PCR efficiency is usually caused by the presence of non-specific/primer dimer product registering fluorescence or PCR inhibitor. Nonspecific amplification could be the obvious cause for too high efficient in *HDAC4*. However, this too high efficiency was still observed in *ACTB* assay (with no non-specific product). This suggest that double-strand DNA in the reaction that is not detectable by agarose gel electrophoresis or the melting curve analysis may still register fluorescence.

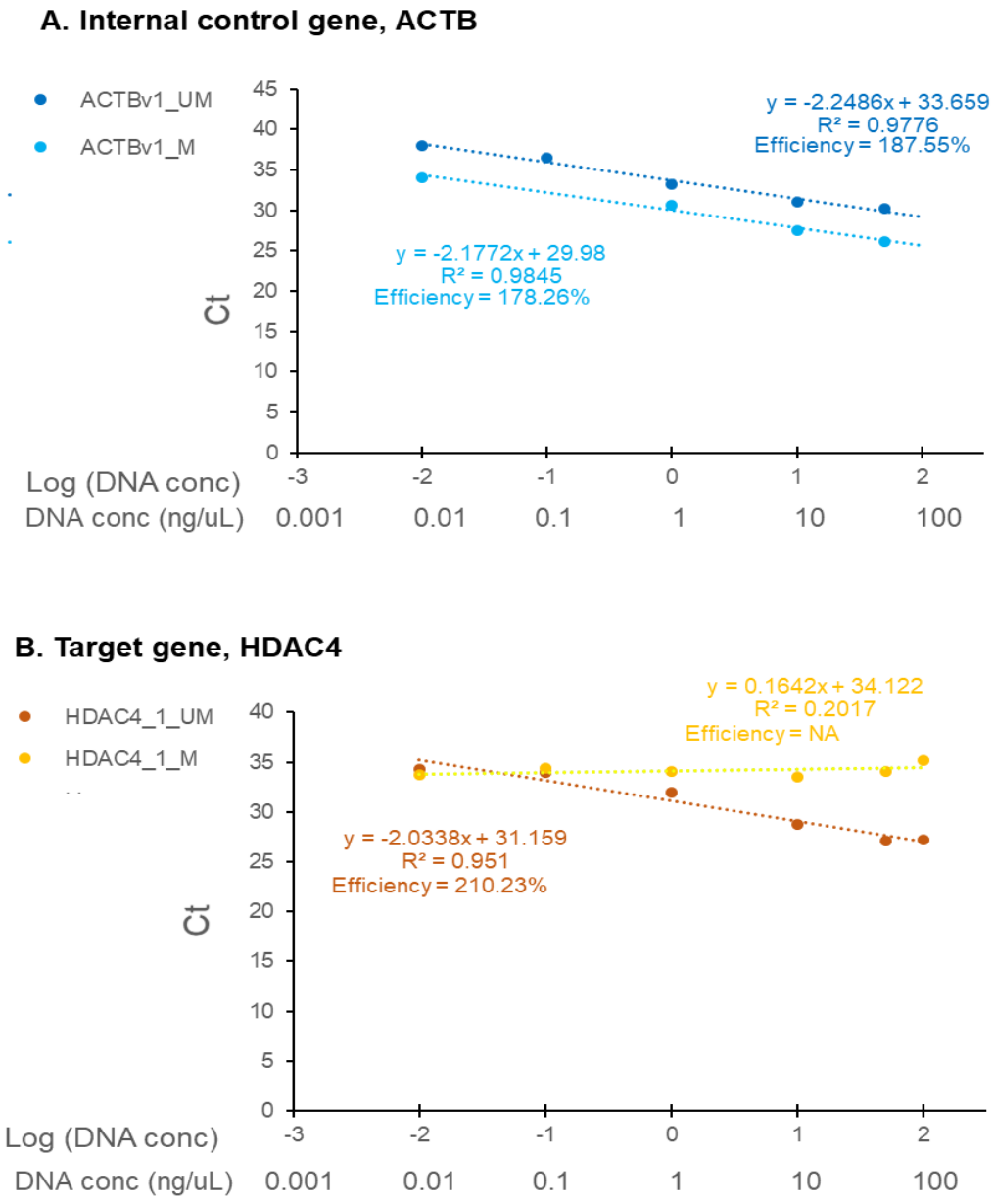


Figure 5-17 Standard curve shows PCR reaction efficiency of A) internal control gene; *ACTB* and B) target gene; *HDAC4* using DNA control as template. UM (100% unmethylated control DNA), M (100% methylated control DNA), NTC (no template control).

Technical discussion

All SYBR green assay optimisation results including primer validation, assay specificity test and assay efficiency test of all primer sets are summarised in Table 5-5. Overall results for developing of qMSP assay using SYBR green-based detection method were disappointing. SYBR green-based detection is usually performant, cost-effective and relatively easy to design and set up, relaying only on good forward and reverse primers design for different assays. However, in this particular case, the SYBR green primer design did not provide sufficient specificity. This dye binding to any double-stranded DNA formed during PCR (including primer-dimer, and nonspecific product) also generated false-positive signals. Well-designed primers and assay optimisation to avoid primer dimer or amplifying of the non-target sequence were therefore a limiting factor in this strategy due to the reduced complexity in the DNA sequence following bisulfite conversion.

Altogether only 4 genes (*HDAC4, TNF, MIR21, RPTOR*) were allowing primers design (out of the 8 top candidate CpGs selected form 3 strategies) while other candidate genes had inadequate nucleotide sequence for primer design or have a limited number of neighbouring CpG to ensure specificity of the primer to un-methylated DNA. Some candidate CpG were also located in the area having (i) a long run of identical nucleotides (i.e. *PSMB9*), (ii) a very low GC content causing very low primer tm, (iii) repeat nucleotide pair that like to form a secondary structure. Some candidate CpGs were an isolated CpG i.e. *ATP6V1H*.

One CpG site in primer was not enough to ensure specific amplification discriminating between un-methylated and methylated sequences. In this experiment, It was very clear that a number of CpG on forward and reverse primer highly affects the reaction specificity (summary Table 5-5) and the, only assay discriminating UM and M-DNA was that which contain 6 CpG in the primer set (*HDAC4*). An attempt to increase reaction specificity such as increase the annealing temperature did not seem to help.

For all these reasons added to the issue of over 100% efficiency of the PCR reaction and time pressure, I decided to stop developing SYBR green-based detection qPCR assays and to switch to Tag-man based detection.

			Primer validation			Assay specificity	Results	
	Genes	**Total number of CG in F+R**	**Temp(°C)**	**F/R primer conc (ng)**	**Primer dimer**	**Metylation Specificity**	**ΔCt (UM/M)**	**NTC**
Control genes	**Ideal reaction**		**60**	**F/R**	**no**	**Equal amplification UM / M**	**none**	**no amp / 40**
	GAPDH V1	0	**60**	**F50/R50**	**yes**	**nearly equal**		detected
	GAPDH V2	0	60	F50/R50	**yes**	equally amp		detected
	ACTB V1	0	**60**	**F50/R200**	no	**equally amp**	<1	none
	ACTB V2	0	60	F900/R50	no	nearly equal		none
	ACTB V3	0	61	F50/R300	no	nearly equal		none
Target genes	**Ideal reaction**		**60**	**F/R**	**no**	**Specific for UM only**	**>6 cycles**	**no amp / 40**
	HDAC4 V1	6	**60**	**F50/R50**	**yes**	**Specific for UM**	~4	detected
	HDAC4 V2	4	60	F900/R900	no	no	~5	none
	HDAC4 V3	4	60	F50/R50	no	no	~5	none
	TNF V1	5	**60**	**F50/R300**	no	**no (seems lilkely)**	~3	none
	TNF V2	2	60	F50/R300	no	no	<1	detected
	MIR21	1	60	F300/R50	no	no	<1	detected
	RPTOR	2	60	F300/R50	no	no	no amp	none

Table 5-5 Summary of SYBR green qMSP assay optimisation result including primer validation, assay specificity test of 12 primer sets. The ideal reaction of internal control gene and target genes were also illustrated.

5.3.3.2 Optimisation qMSP condition for target and internal control genes using TagMan based assay

qPCR using TaqMan-based detection may help improve the reaction specificity that occurred with SYBR green based assay by introducing a probe tagged with fluorescence. This ensures that the fluorescent signal comes truly from the target product (i.e. no issue with primer dimer or non-specific product). Introduction of the probe also allows more CpG sites in the assay helping to increase discriminating power of between un-methylated and methylate DNA.

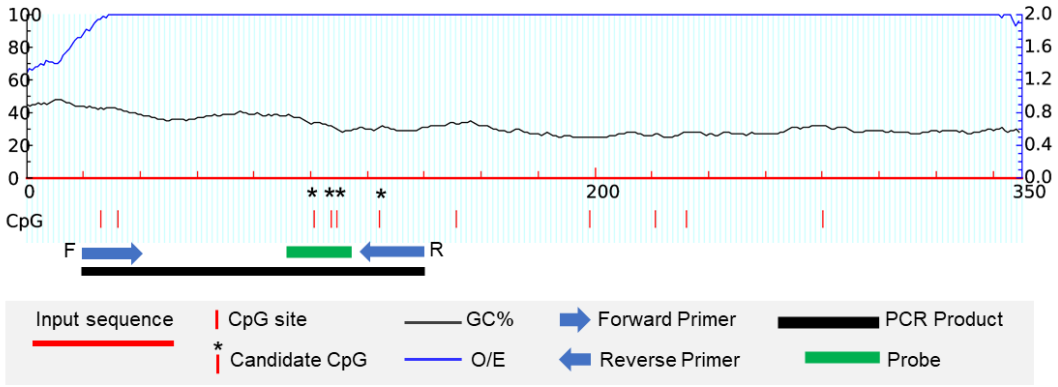
Again, I designed 2 types of the assay, one for target genes and one for the control gene used to normalise the amount of template DNA input. Throughout the process, 100% methylated or/and 100% unmethylated DNA control were used as an input DNA template.

The assay development followed the same workflow. First, beginning with primers design and primer validation to find the right concentration of primers that give the highest fluorescent intensity. Then checked primer specificity to methylated DNA for the target gene assay, and the methylation independent amplification for control assay, then the final step comparing the reaction efficiency. However, I also made the decision to design assays for detecting methylated DNA for the target gene assay to increase GC content in primers and probes.

5.3.3.2.1 Primers and probes design

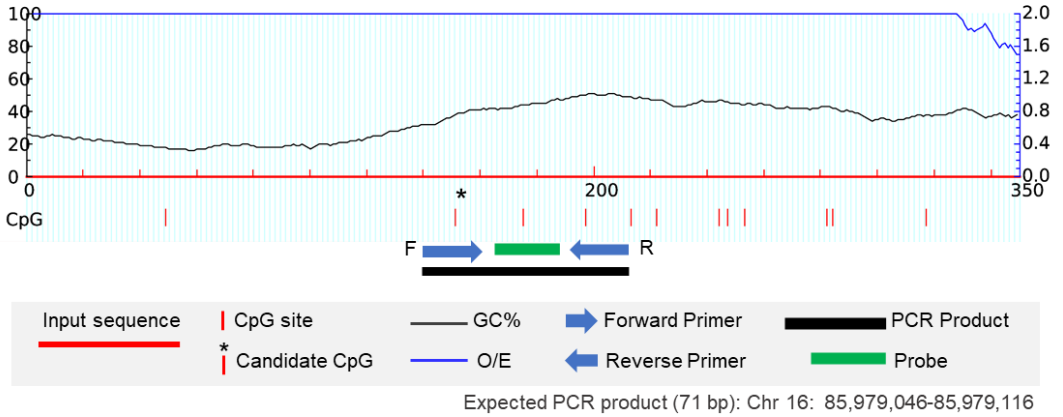
I aim to design 3 target genes assay and 2 internal control gene assay. Schematic diagram of primer and probe design of each genes and the position of the candidate CpG presents in Figure 5-18. The sequences and other details of all primers were listed in the Appendix 5.

A TNF



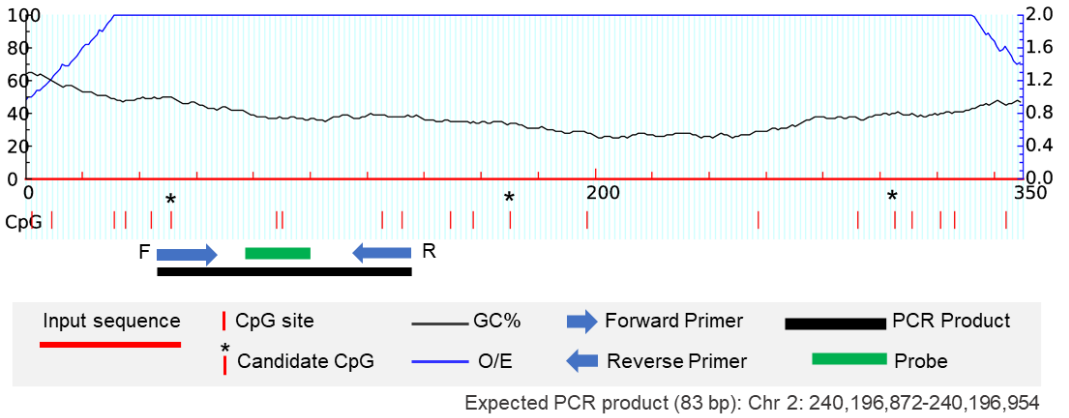
Expected PCR product (121 bp): Chr 6: 31,543,091-31,543,211

B IRF8



Expected PCR product (71 bp): Chr 16: 85,979,046-85,979,116

C HDAC4



Expected PCR product (83 bp): Chr 2: 240,196,872-240,196,954

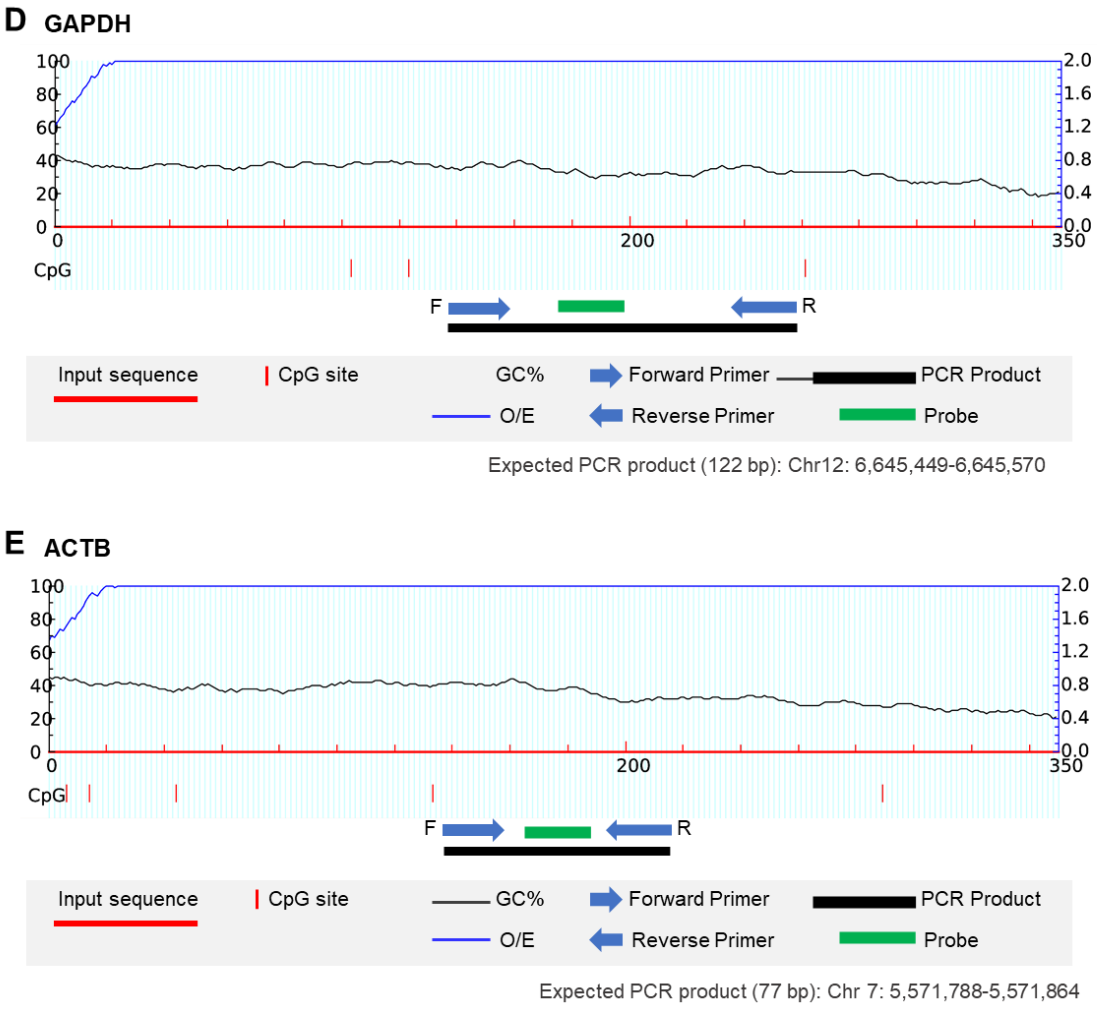


Figure 5-18 Schematic diagram of primer and probe design of the target genes; A) *TNF*, B) *IRF8*, C) *HDAC4* and the internal control genes; D) *GAPDH*, E) *ACTB*.

5.3.3.2.2 Optimisation of PCR condition

Primer validation

Primer validation was aiming to find the right concentration of forward and reverse primers that give the best Ct (yield the lowest Ct). 9 primer concentration of forward and reverse primers combination varied from 50-900 nM were tested using 10 ng of 100% methylated DNA control.

For *TNF* for example, all combination of primer concentration gave very similar results (Figure 5-19) while for *GAPDH* gene, low concertation of R-primer had a clear effect on the amplification independently of the amount of F-primer (Figure 5-20).

The final primer conditions that gave the best Ct values were chosen at F900/R900 nM for *TNF, HDAC4*, and *ACTB* assay, F300/R900 for *GAPDH*, and F300/R300 for *IRF8*.

The specificity of the assay for methylated DNA

For target gene assay, the specificity for methylated CpG was preliminarily observed by comparing the amplification result of 100% methylated DNA control to 100% un-methylated DNA control. The specificity for the methylated sequence was confirmed for *TNF* and *HDAC4* assay as showed by good amplification (Ct =28.97± 0.02, and Ct=29.72 ± 0.08) for 10 ng methylated DNA control with no amplification at all for 10 ng un-methylated DNA template (Figure 5-19, blue line). These assays were future tested for PCR efficiency.

On the other hand, the *IRF8* assay was not totally specific for methylated DNA and showed amplification with the un-methylated template. Primer concentration at F300/R300 mM which provided the lowest Ct value (Ct= 29.58±0.16) for the methylated DNA showed a large cycle different between of two types of DNA template (~11 cycles difference) (Figure 5-21). This suggest a possible amplification of a 2^{-11} magnitude which may be considered neglectable. These conditions were therefore used for further optimisation steps.

The methylation independence for *GAPDH* and *ACTB* assay was confirmed by the equal amplification using methylated and un-methylated DNA (data in Appendix 10).

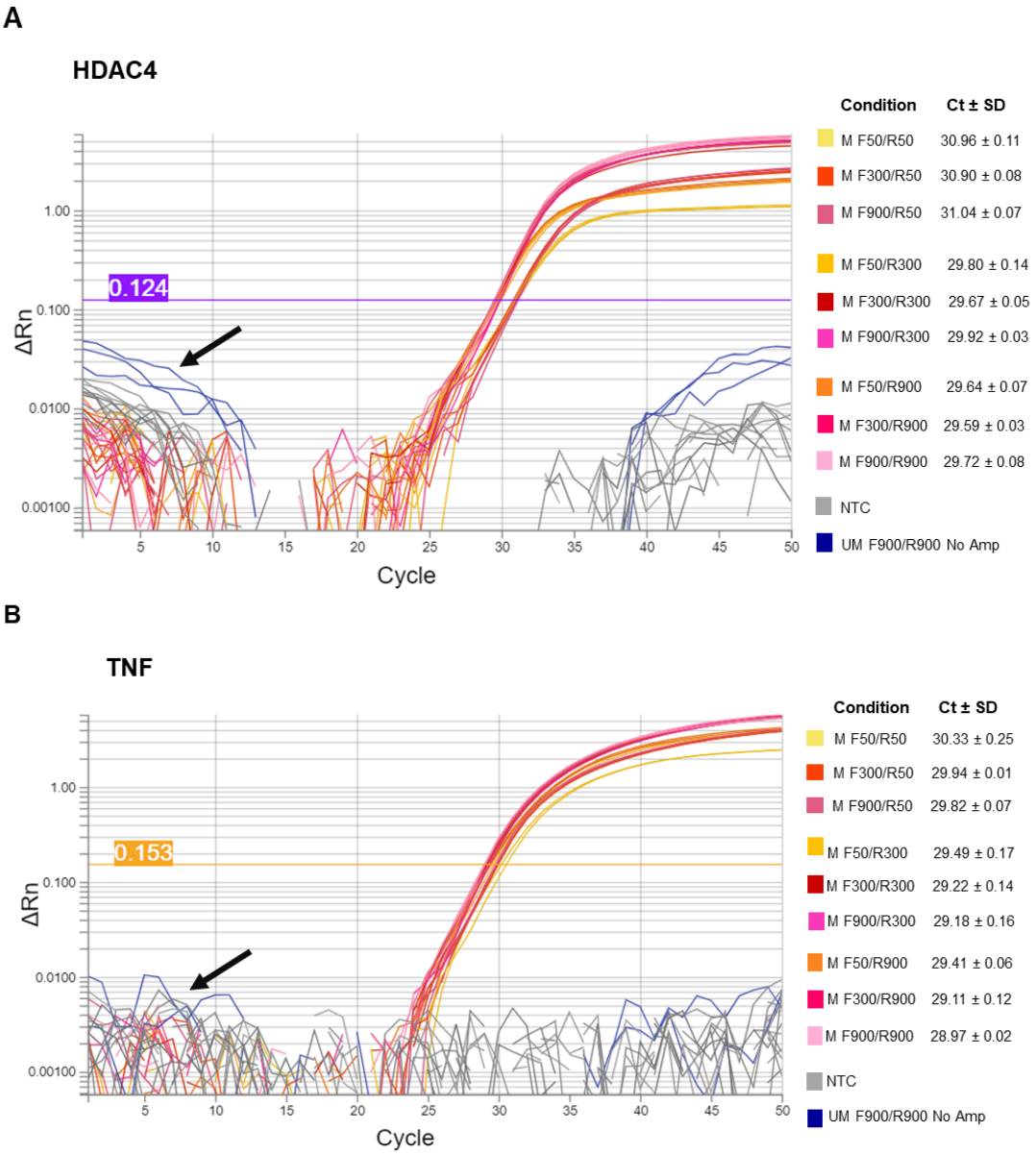


Figure 5-19 Amplification plot of A) *HDAC4* and B) *TNF* assay using forward and reverse primer at different concentration (nM). M: 100% methylated control DNA, UM: 100% un-methylated control DNA (highlighted by an arrow), NTC: no template control. For both assays, the primer concentration at F900/R900 nM was chosen for the further test. The assay specific to the methylated sequence was confirmed by no amplification of the un-methylated template.

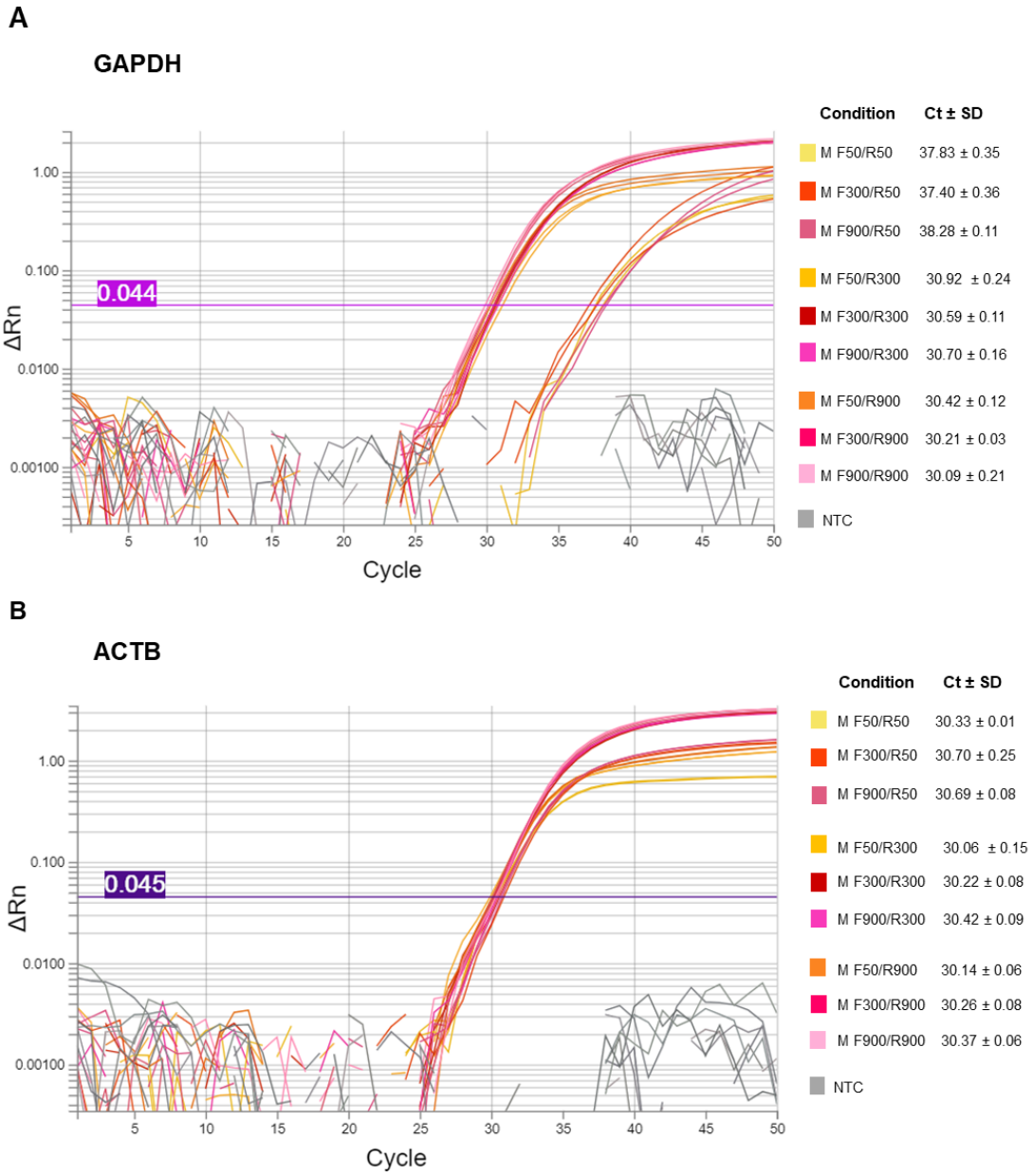


Figure 5-20 Amplification plot of A) *GAPDH* and B) *ACTB* assay using forward and reverse primer at different concentration (nM). M: 100% methylated control DNA, UM: 100% unmethylated control DNA, NTC: no template control. The primer concentration at F300/R900 and F900/F900 nM was chosen for the further test for *GAPDH* and *ACTB* assay, respectively.

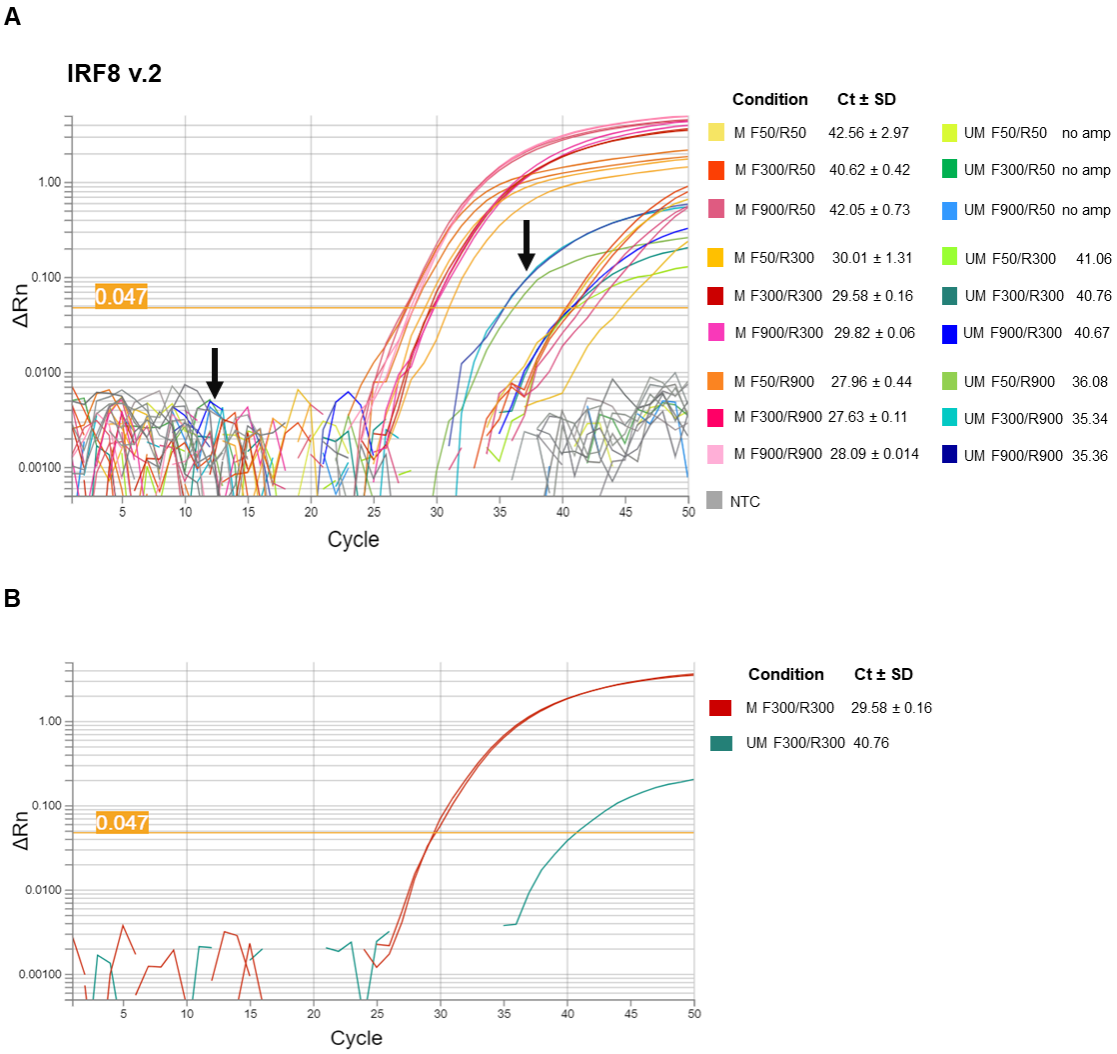


Figure 5-21 Amplification plot of *IRF8* assay A) using forward and reverse primer at different concentration (nM). M: 100% methylated control DNA, UM: 100% un-methylated DNA control (an arrow), NTC: no template control. The present of un-methylated DNA amplification suggested no specific to methylated DNA of *IRF8* assay B) using methylated and un-methylated DNA at best primer concentration, F300/R300 mM. Nonspecific to the methylated sequence was observed at 11 cycles slower.

Varying annealing temperature

To improve the *IRF8* assay, I also relied to optimise the qPCR by varying the annealing temperature ranging from 58 °C to 64°C. Increase temperature from 60 °C to 61 °C helped improve the reaction specificity to methylated DNA (adding 1.6 more cycle difference between un- and methylated DNA) (Figure 5-22). Although increasing temperature improves the Ct-different it ,however, increases the actual Ct value from 29 to 34 which is too high. The slope of the amplification plot was also less linear suggesting lower efficiency.

A two-step annealing, performing the PCR at high temperature for 5 sec, follows by the normal temperature at 60°C for 55 sec were tested. Using this technique helped maintain the Ct value for methylated DNA at <30 but did not stop amplifying un-methylated DNA with 9 cycle (61°C, 5 sec and 60 °C, 55 sec condition) difference between methylated and un-methylated DNA.

Overall the best condition for *IRF8* assay selected for further step used F300/R300 nM, PCR annealing/extension temperature at 62°C,5 sec then 60°C, 55 sec. This gives Ct of 29.58 and 12 cycle difference between M and UM DNA.

.

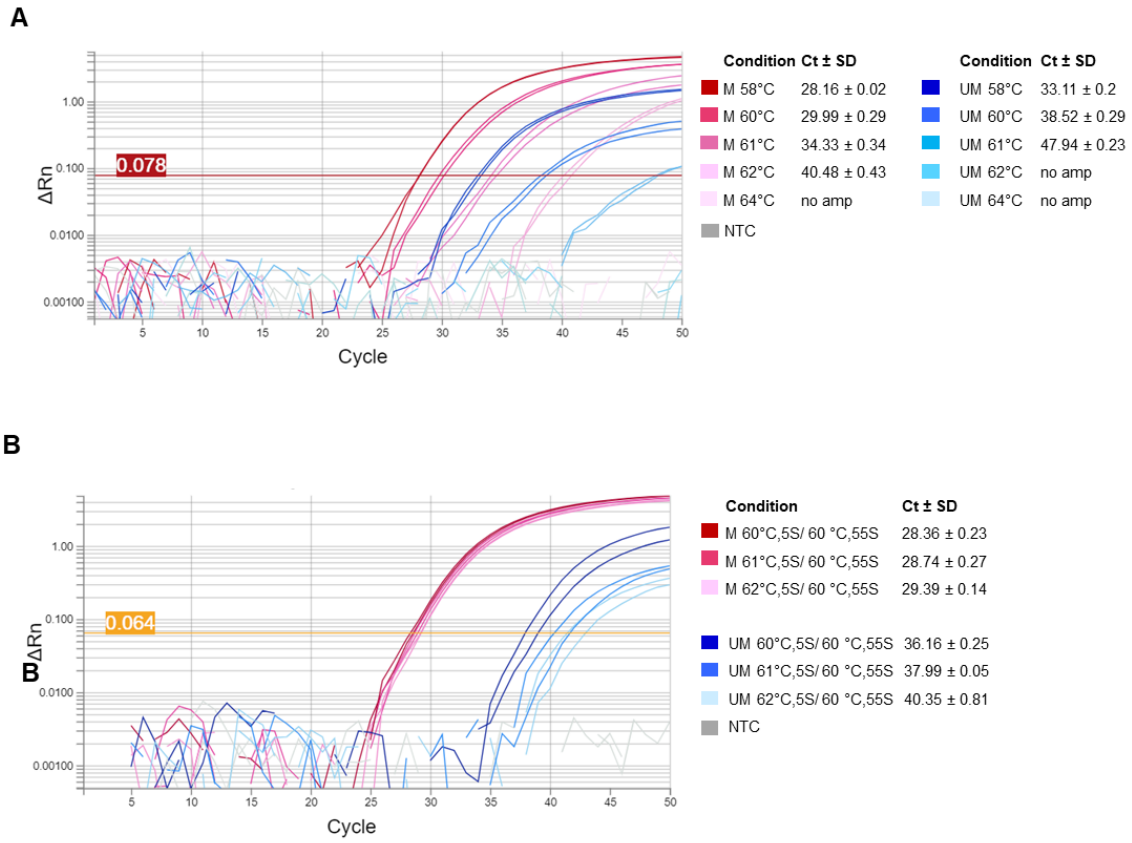


Figure 5-22 Amplification plot of *IRF8* assay at different annealing temperature. M: 100% methylated control DNA, UM: 100% unmethylated control DNA, NTC: no template control.

5.3.3.2.3 TagMan qPCR efficiency

qPCR efficiency of target gene assay (*HDAC4, TNF*, and *IRF8*) and internal control assay (*GAPDH* and *ACTB*) were accessed using a dilution curve. The amplification of serial dilution of 100% methylated and un-methylated DNA controls at a concentration range from 0.2 ng to 50 ng was performed (amplification plot in appendix 11). The Ct value from each condition of the individual assay was obtained and plotted against Log (DNA concentration (ng/ µL)) (Figure 5-23 for *GAPDH, ACTB, HDAC4, TNF*, and Figure 5-24 for *IRF8*).

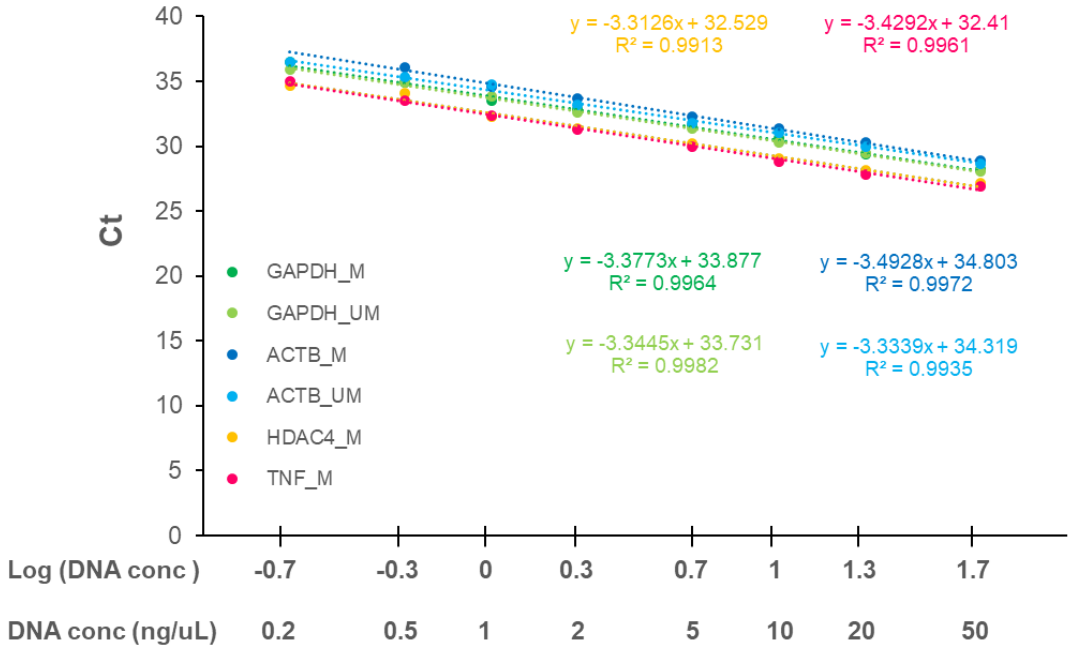
Fitting the standard curve to linear regression model provides the slope used to calculate PCR efficiency (Figure 5-23). The regression showed true linearity over the whole range of concentration suggesting that the amount of DNA input could be quite variable (as extracted from real patient's sample) without composing the result. A Good range of efficiency was observed (90-110% being optimal) from four assays, *GAPDH* (M:97.74%, UM: 99.07%), *ACTB*(M: 93.33%, UM: 99.50%), *HDAC4* (M: 100.39%, UM: not amplified) and *TNF* (M: 95.71%, UM: not amplified), suggesting a good 2-fold amplification of the amount of DNA at every cycle and good reliability of the assay. R^2 also very close to 1 suggested that the model fit well with data.

For the internal control gene, methylated and un-methylated DNA were equally amplified and also showed a similar efficiency. The efficiency difference between un- and methylated DNA were less for *GAPDH* (1.33%) compared to *ACTB* (6.17%) assay. As a result I chose to use *GAPDH* as my internal control for further work with patient samples.

For the target gene, *HDAC4* and *TNF* assays were proven specific for the methylated sequence at all DNA concentration. The overall Cts were quite close but not totally overlapping for *TNF* with those for *GAPDH* and *HDAC4* for 100% methylated DNA, suggesting small adjustment would be needed when testing real samples.

However, for *IRF8* assay (Figure 5-24), the reaction was less specific to the methylated sequence as expected, but the amplification still showed good linearity on methylated DNA with a stable difference with un-methylated DNA although maybe on a slightly reduced range from 0.5-20 ng of DNA. The reaction efficiency was however outside of the acceptable range (M:124.45%, UM: 83.33%). Therefore, I decided not to use this *IRF8* assay, and to concentrate on *HDAC4* and *TNF* for the work with patient samples.

A



B

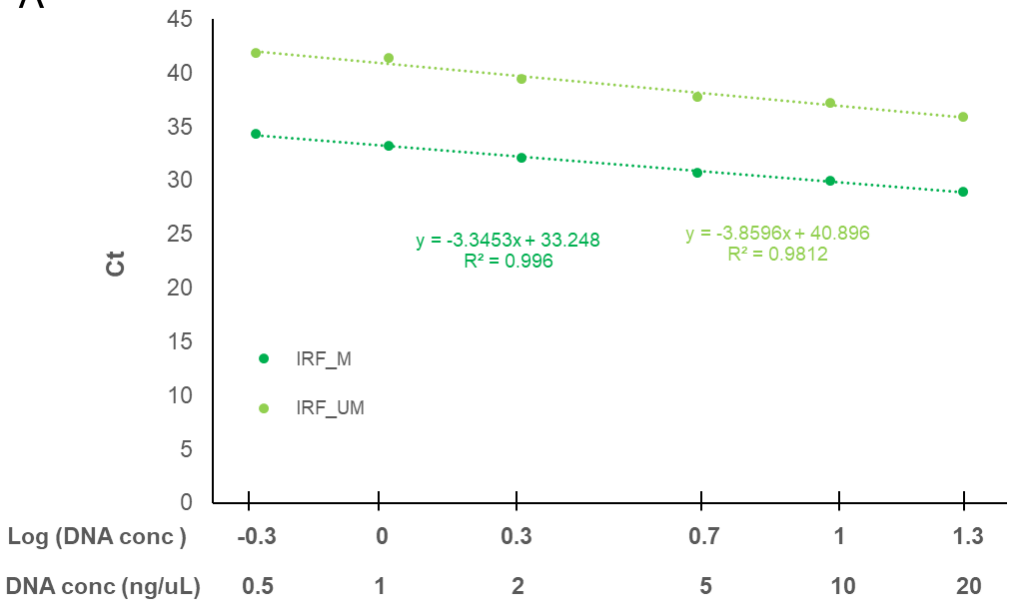
DNA (ng/uL)	Log10 (DNA ng/uL)	GAPDH v10		ACTB v4		HDAC4 V1		TNF v4 (seq)	
		M	UM	M	UM	M	UM	M	UM
50.00	1.70	28.23	28.09	28.89	28.68	27.13	No amplification	26.92	No amplification
20.00	1.30	29.38	29.44	30.29	29.95	28.13		27.79	
10.00	1.00	30.59	30.27	31.41	31.08	29.06		28.84	
5.00	0.70	31.59	31.34	32.24	31.75	30.19		29.95	
2.00	0.30	32.85	32.62	33.68	33.15	31.39		31.26	
1.00	0.00	33.54	33.85	34.63	34.78	32.32		32.33	
0.50	-0.30	34.89	34.92	36.07	35.36	34.06		33.51	
0.20	-0.70	36.45	35.95		36.47	34.70		34.98	
	R^2	**0.996**	**0.998**	**0.997**	**0.994**	**0.991**		**0.996**	
	Slope	-3.377	-3.345	-3.193	-3.334	-3.313		-3.429	
	Efficiency (%)	**97.74%**	**99.07%**	**93.33**	**99.50%**	**100.39%**		**95.71%**	

Figure 5-23 PCR efficiency of 4 assays

A) Dilution curve and linear regression trend line of 4 assays; *GAPDH* (green), *ACTB* (blue), *HDAC4* (orange) and *TNF* (pink). All assay show a good efficiency and good fit of the regression model to the data.

B) Table descript Ct value from different assay using 100% methylated and unmethylated DNA control at different concentration and information form liner regression model (R2, slope) and PCR efficiency of each assay.

A



B

DNA (ng/uL)	Log10 (DNA ng/uL)	IRF8 v2	
		M	UM
50.00	1.70	29.13	37.29
20.00	1.30	29.00	35.96
10.00	1.00	29.98	37.22
5.00	0.70	30.70	37.82
2.00	0.30	32.13	39.47
1.00	0.00	33.28	41.40
0.50	-0.30	34.36	41.92
0.20	-0.70	35.18	46.31
	R^2	**0.097**	**0.084**
	Slope	-2.848	-3.799
	Efficiency (%)	**124.45%**	**83.33%**

Figure 5-24 PCR efficiency of *IRF8* assays

A) Dilution curve and linear regression trend line of *IRF8*. The assay shows a poor efficiency.

B) Table descript Ct value from different assay using 100% methylated and unmethylated DNA control at different concentration and information form liner regression model (R2, slope) and PCR efficiency of each assay.

5.3.3.3 Optimised TagMan qPCR assay

The summary of TaqMan qMSP assay optimisation result including primer validation, assay specificity test and assay efficiency of all assay is described in Table 5-6. The ideal reaction requirement for Target gene assay and internal control assay also described.

The successful design assay that will be used to perform on patient samples were *HDAC4, TNF*, and *GAPDH*. The qMSP reaction composition and cycling condition displayed in Table 5-7.

TagMan qMSP technical discussion

As mention before, not all candidate genes have the potential to be used for qMSP assay design. *PSMB9* have an inadequate number of neighbour CpG around the candidate CpG. *PRTOR, ATP6V1H* and *MIR21* candidate CpG are located in an area that are not facilitating primer binding. Because of these limitations and also time restrictions, the candidates with lower potential or more complicated design could not be developed.

The TagMan based detection method increases the specificity of PCR by the introduction of fluorogenic-labelled probes allowing an added layer for the target-specific signal detection compared to SYBR green assay which detects (any)double-strand nucleotide.

4 assays; *TNF, HDAC4, GAPDH,* and *ACTB* were successfully optimised while the *IRF8* assay was not optimal most likely due to the lack of sufficient specificity/number of CpG sites in the methylated sequence. Optimisation by increasing the annealing temperature to improve the reaction specificity had limited usefulness. Experiments using another *IRF8* primer set which included one more CpG was also tested but show no improvement (detailed in Appendix11, *IRF8* V.2).

Table 5-6 Summary of TaqMan qMSP assay optimisation result including primer validation, assay specificity test and assay efficiency test of 6 primer sets. The ideal reaction of internal control gene and target genes were also illustrated.

			Primer validation			Assay specificity	Results	
	Genes	Total number of CG in F+R+P	Temp(°C)	F/R primer conc (ng)	Probe conc (ng)	Metylation Specificity	ΔCt (UM/M)	NTC
Control genes	**Ideal reaction**		**60**	**F/R**	**P**	**Equal amplification UM / M**	**none**	**no amp**
	GAPDH	0	**60**	**F300/R900**	**250**	**equally amp UM and M**	<1	no amp
	ACTB	0	60	F900/R900	**250**	equally amp UM and M		
Target genes	**Ideal reaction**		**60**	**F/R**	**no**	**Specific for UM only**	**>6 cycles**	**no amp**
	HDAC4	4	**60**	**F900/R900**	**250**	**Specific for M**	no amp for UM	no amp
	TNF	**6**	**60**	**F900/R900**	**250**	**Specific for M**	no amp for UM	
	IRF8	3	62 5s,60 55s	F300/R300	250	no	~10	
	IRF8 V2	4	60	F50/R50	250	no	~8	

Table 5-7 Optimised Tag-man qMSP reaction composition and cycling condition.

qPCR TagMan assay

Stock conc.	Reagent	Final concentration		
		GAPDH	HDAC4	TNF
2X	universal mastermixII no UNG	1X	1X	1X
vary	F primer	300 nM	900 nM	900 nM
vary	R primer	900 nM	900 nM	900 nM
2.5 uM	Taqman probe	250 nM	250 nM	250 nM
	DNA template	20 ng/uL	20 ng/uL	20 ng/uL

Total Volumn 20 ul

PCR cycling condition

Step	Temperature	Time	
Initial activation	95 C	10 min	
Denaturation	95 C	15 s	50 cycles
Annealing/Extension	60 C	60 s	

Similar to the SYBR green assay discussion, several factors are involved in the success of qMSP reaction. The nature of the lower DNA complexity after bisulfite conversion is a major challenge . A dinucleotide repeats or long run of identical nucleotides near the candidate CpG cause the difficulty in designing good primer. The low CG content also affects primer binding and tm. In qMSP assay, primers could be designed to detect either methylated or un-methylated sequence. By choosing to design assay on methylated DNA, the use of CG instead of TG bases in primer sequence increase primer tm making the reaction more specific. Thus, designing primer to detect methylated sequence proved a good choice as it resulted in better assay performances.

The TagMan qPCR technology either the one for gene expression assay or qMSP assay was known for its high-throughput quantification, high sensitivity, and reproducibility(259, 382). The TagMan qMSP assay were showed to accurately determine the relative prevalence of a particular pattern of DNA methylation.

For quantification, the methylation level in samples relies on the relative quantification method with normalisation to an internal control gene. It is also important to ensure that the assay efficiency of all genes are equal (or nearly equal) and within the acceptable range of 90-110%.

For the sensitivity of the assay, the dilution curve analysis confirmed the good working range of DNA input between 0.2 ng to 50 ng of DNA. For the patient samples with unknown methylation level, I therefore decided to use 20 ng considering that the detection of 0.2 ng was still in linear range and is theoretically equivalent to 1% methylation in 20 ng DNA. This amount of DNA input is also in line with others published assays (10 to 100 ng) (260, 287, 383, 384).

In the next step, the well-optimised qMSP assay; *TNF* and *HDAC4* were tested in patient samples to determine whether they can be used as a diagnostic biomarker for RA or they used as a marker to predict the response to the treatment.

5.3.4 Performance of the qMSP assays in RA and Non-RA classification performed on patients samples

To prove whether methylation of the *HDAC4* and *TNF* promotor at the candidate loci can be used as biomarkers for the classification of RA, the optimised qMSP assays were tested on patient samples with different diagnoses including RA, versus non- RA arthritis (e.g., possibly UA, PSA, reactive arthritis, connective tissue or gout), as well as in healthy donors.

The experimental design and analysis are described in the flow chart below.

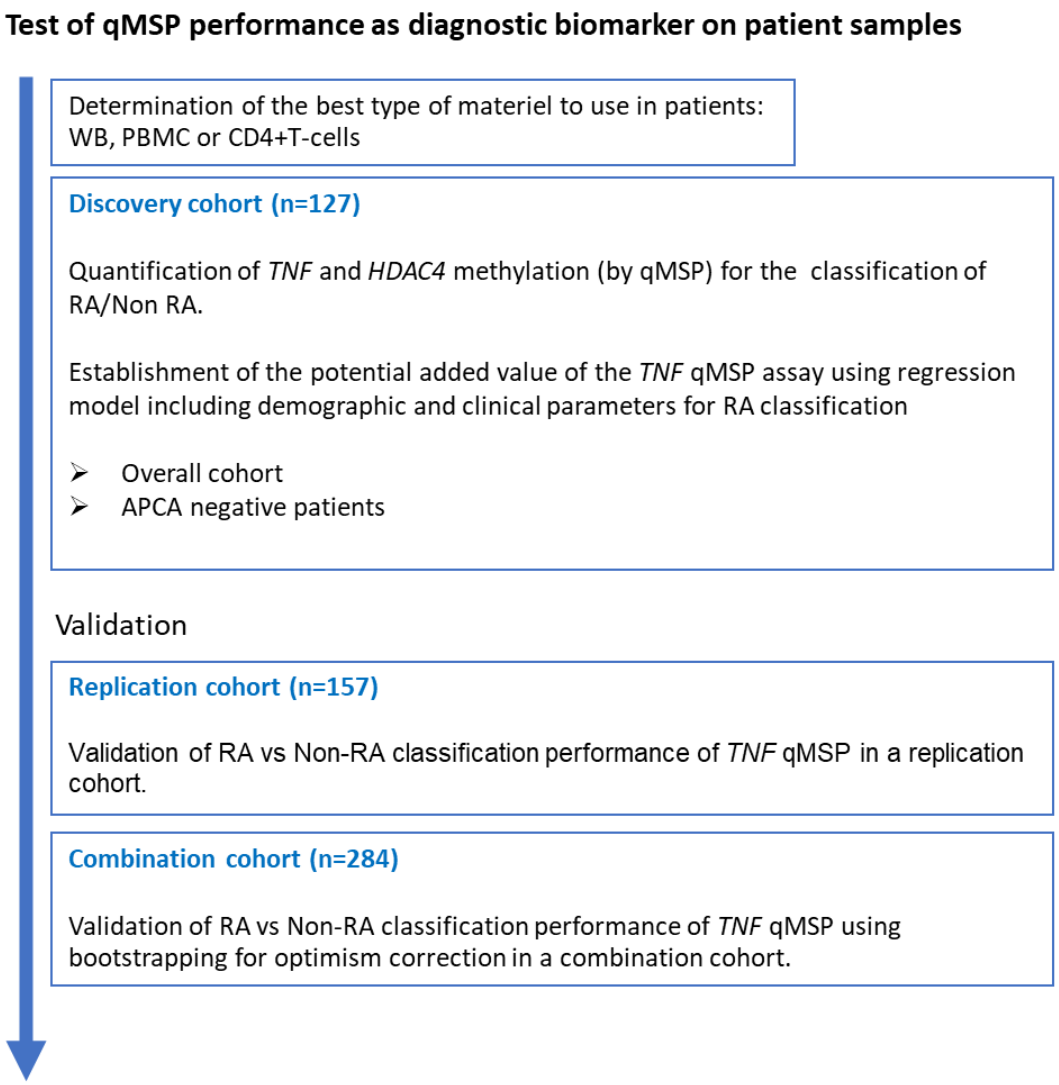


Figure 5-25 Flow chart for testing the qMSP assay performance as diagnostic biomarker

5.3.4.1 Determination of the best type of materiel to use in patients : WB, PBMC or CD4+T-cells

For biomarker development, the design of the biomarker assay would benefit from the least processed sample, in order to limit variability introduced during additional processing steps. To decide which type of samples is best suited to use for a quantitative assay, different samples types: CD4+T-cell, PBMC and WB were preliminarily tested in a small group of patient with a different diagnosis (HC, RA and UA). Samples could not be matched for the same individual, however, they were selected based on a similar age range. All samples were part of the IACON cohort. WB DNA had previously been extracted for a similar study (378). Frozen PBMC cells and CD4+T-cells isolated from frozen PBMC by magnetic bead negative selection were also used. The number of sample and the age range for each group of samples for the different diagnosis are described in Table 5-8.

DNA isolated from these cells was bisulfite converted and used in the qMSP assay. For each sample, target gene reactions were run in parallel to the control reaction. 100% methylated, 100% unmethylated DNA controls, and NTC were included in each plate. Methylation level (%) were quantified from the relative to 100% methylated DNA control as described in the method section. % Methylation for both the *TNF* and *HDAC4* qMSP quantification are described in Table 5-8 and illustrated by boxplot in Figure 5-26.

Table 5-8 Description of results for levels of methylation (%) of the *TNF* and *HDAC4* gene determined by qMSP in different clinical groups

A TNF

Diagnosis	Sample type	n	Age		Methylation (%)	
			Mean	SD	Mean	SD
HC	CD4+T-cells	6	39.15	7.27	4.76	0.56
RA	CD4+T-cells	6	44.50	7.50	2.18	0.90
HC	PBMC	6	41.83	5.53	6.59	1.73
RA	PBMC	6	39.67	6.77	3.80	1.35
UA	PBMC	6	41.67	4.23	5.71	1.01
RA	WB	9	55.78	15.79	6.07	2.03
UA	WB	5	43.00	9.57	6.19	1.59

Comparison test	MWU test		
Cell type	Comparison		P value
CD4+T-cells	HC	RA	0.0022
PBMC	HC	RA	0.0152
PBMC	HC	UA	0.4848
PBMC	RA	UA	0.0260
WB	RA	UA	0.7972

B HDAC4

Diagnosis	Sample type	n	Age		Methylation (%)	
			Mean	SD	Mean	SD
HC	CD4+T-cells	6	39.15	7.27	16.31	6.34
RA	CD4+T-cells	6	44.50	7.50	8.40	3.26
HC	PBMC	6	41.83	5.53	15.21	5.40
RA	PBMC	6	39.67	6.77	14.85	4.70
UA	PBMC	6	41.67	4.23	13.53	2.05

Comparison test	MWU test		
Cell type	Comparison		P value
CD4+T-cells	HC	RA	0.0260
PBMC	HC	RA	1.0000
PBMC	HC	UA	0.8180
PBMC	RA	UA	0.9370

Statistical analyses comparing clinical groups were performed using MWU test or the Kruskal-Wallis test followed by Dunn's multiple comparison test, for the *TNF* and *HDAC* methylation assays for each cell type. Results were not corrected as only used for establishing best type of samples to use.

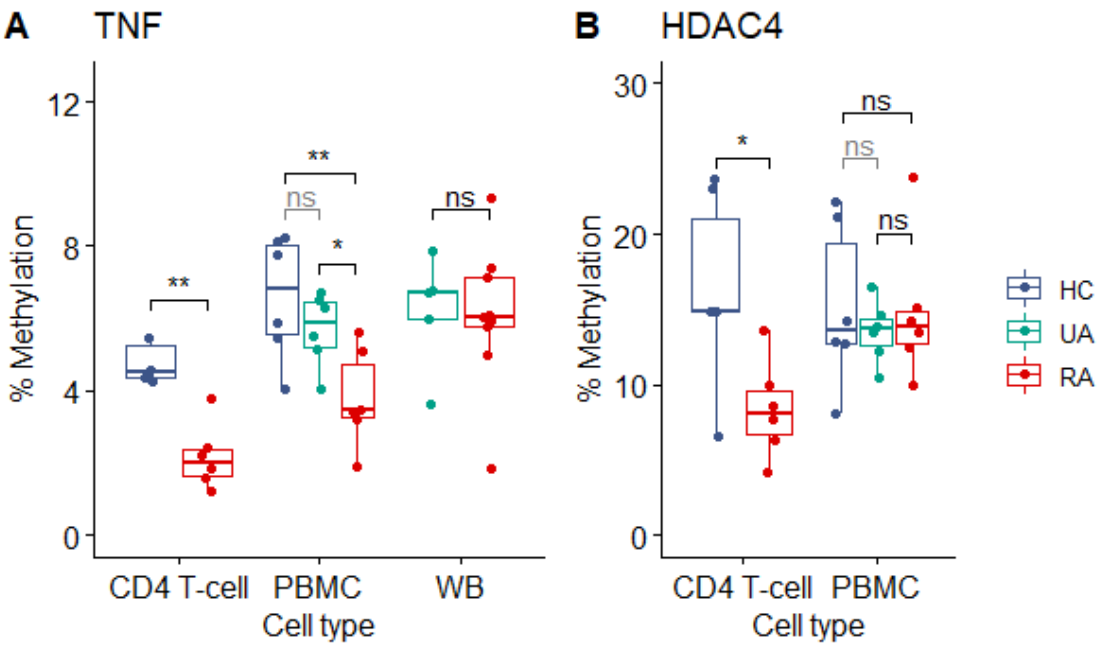


Figure 5-26 Boxplot of the levels of methylation (%) in CD4+T-cells, PBMC and WB in different clinical groups for the *TNF* (A) and *HDAC4* (B) assays. Statistical analyses were performed using MWU test (2 groups) or the Kruskal-Wallis test followed by Dunn's multiple comparison test (3 groups), (**p<0.001, *p<0.05, ns non-significant).

For TNF qMSP assay

The methylation levels (mean±SD) in HC CD4+T-cells and PBMC were 4.76 ± 0.56% and 6.59 ± 1.73%, respectively; in RA they were 2.18 ± 0.90% and 3.80 ± 1.35%. Of note, the methylation levels quantified by the qMSP assay showed lower readings compared to the methylation levels (%) detected in publicly available Illumina methylation wide array dataset (~40% for CD4+T-cells and ~45% for PBMC in HC) or our bisultfie sequencing data (~45% for CD4+T-cells in HC). This difference in % of methylation detected between the 2 techniques will future discuss in the discussion part.

Nevertheless, The *TNF* qMSP methylation levels in RA were clearly lower than in HC in CD4+T-cell (2.2 fold reduction in median/mean, p=0.002) as well as in PBMC (1.7 fold, p= 0.0152). The higher fold-reduction in methylation in CD4+T-cells suggested that DM in this locus is highly T-cells specific. The lower fold-reduction in PBMC is likely to results from a “dilution” effect on the overall methylation levels, as CD4+ T-cells represent about >50% of PBMC.

In PBMC, DM between RA and UA (3.80 ± 1.35% and 5.71 ± 1.01%) was observed, with a 1.5 fold reduction (p=0.026) while no significant DM was detected between HC and UA. This suggesting that the TNF methylation in PBMC can differentiate between RA and non-arthritis group thus has the potential to be used as a diagnostic biomarker which would be easier access source of material than purified CD4+T-cells.

In WB, The difference in methylation between RA (mean 6.07 ± 2.03%) and UA (6.19 ± 1.59%) samples was not statistically significant (p=0.797).

For HDAC4 genes,

Methylation levels in RA were lower than HC in CD4+T-cells (p = 0.026), however no difference in PBMC. DM between HC and RA was clear in CD4+ T-cells (1.94 fold reduction in CD4+T-cells) suggesting that DM in *HDAC4* is also highly CD4+T-cells specific. The lack of DM using PBMC suggests that the signal was highly diluted by other cells in PBMC (compared to the *TNF* gene signal notably). This was an important observation, suggesting that in RA the lower methylation of the *TNF* gene in PBMC DNA may also results from lower methylation in CD8+T-cells, B-cells or NK cells in RA although data publicly available however in established RA are not suggesting significant DM in any individual cells types.

There was also no change in *HDAC* methylation levels between RA compared to UA in PBMC samples. The *HDAC4* qMSP assay cannot differentiate RA group from UA or even RA from HC. This suggest that HDAC is not likely to be a have value as a biomarker of RA diagnosis, although it confirms that DM is important as well as CD4+T-cell specific from a disease mechanism point of view.

These results altogether, suggested that WB cannot be used for a diagnostic biomarker test, as the dilution effect of having DNA from non-CD4+T-cells in the samples was limiting the ability to detect changes. Therefore, PBMC which is the next least processed type of sample allowing discrimination was chosen for further use to study the value of the *TNF* qMSP assay as a diagnosis biomarker.

5.3.4.2 Discovery cohort for determining the potential of the qMSP assay as diagnosis biomarker

DM of the *TNF* gene promoter between RA and other arthritis using qMSP assays

I then examined whether the *TNF* (and *HDAC4*) qMSP assay had the potential to be used as a diagnostic biomarker, using PBMC DNA of patients with RA, other types of arthritis (non-RA), as well as healthy individuals.

A total of 158 frozen PBMC from 65 RA, 64 non-RA (including 11 reactive arthritis, 37 undifferentiated arthritis (UA), and 16 psoriatic arthritis (PSA)), and 29 HC was obtained from our tissue bank. DNA was isolated, bisulfite converted, and used as template for the *TNF* qMSP assay as described earlier.

This group was selected from the IACON register and were collected recruited between 2010-2014. It will be further referred as to the IACON Discovery cohort.

For the TNF assay (Figure 5-27*,A, Table* 5-9*,A)*

Two samples which had unreliable TNF qMSP value due to poor quality DNA were removed from further analysis leaving 29 HC, 64 RA and 63 Non-RA samples (11 reactive arthritis, 36 UA, 16 PSA). Methylation between each diagnosis group was significantly different ($p=3.5x10^{-9}$). Comparing RA to individual groups, there was a significant DM with HC (reduction of the median % of methylation, Δ-DM= - 3.37% $p=1.97 \times 10^{-5}$), reactive arthritis (Δ-DM= -3.80 %, $p= 4.21x10^{-4}$), UA (Δ-DM= -3.12%, $p= 4.26x10^{-7}$) but not with PSA (p=0.155) despite higher % of methylation in PsA (Δ-DM= - 2.07%). In addition, there was no significant DM between HC and the reactive or UA groups.

A clearer view of DM between RA and the overall non-RA group (including reactive arthritis, UA, and PSA) was also showed in Figure 5-27,B (Δ-DM= -3.02%, $p=4.1x10^{-9}$). This result confirmed that *TNF* qMSP has the ability to segregate RA patient from non-RA arthritis, although less efficiently between RA and PsA. Whether it can be used to predict a diagnostic of RA will further be investigated using logistic regression in the next section.

For the HDAC4 assay (Figure 5-27*,C, Table* 5-9*,B)*

Two samples which had unreliable HDAC4 qMSP value due to poor DNA quality were removed from the further analysis. There were also one missing data leaving the total of 29 HC, 64 RA and 62 Non-RA samples (10 reactive arthritis, 36 UA, 16 PSA). Large SDs in the methylation levels were observed between groups, (especially reactive arthritis). After performing a statistical analysis, methylation between each diagnosis group showed overall significant differences (p=0.0452). Individually, this analysis resulted in a significant, although very small Δ-DM between RA and UA (Δ-DM= -7.07%, p=0.0257) while not significant between RA and reactive group (Δ-DM= -6.8%, p=1.00) or PSA group (Δ-DM= -5.59%, p=1.00). Altogether. DM between RA and non-RA group was almost significant (Δ-DM= -5.59%, p=0.0027). This data suggested that *HDAC4* qMSP assay may be used as diagnosis biomarker, although the test performance are unlikely to provided added value as it only allows segregation of RA from arthritis with a rather low inflammatory profile (UA) but not reactive and psoriatic arthritis.

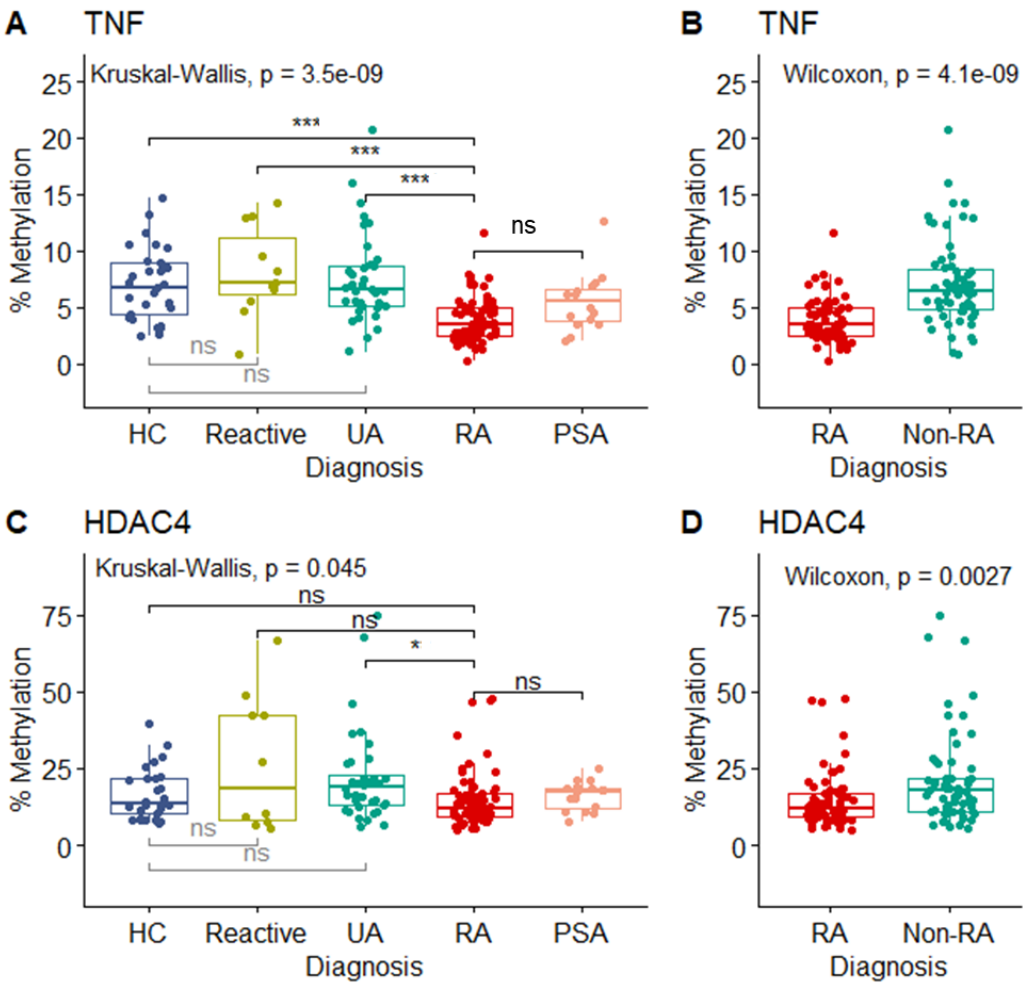


Figure 5-27 Box plot of the levels of DNA methylation (%) for A, B) *TNF* and C, D) *HDAC4* genes using the qMSP assays in different diagnosis groups. Statistical analyses were performed using the Kruskal-Wallis test followed by Dunn's multiple comparison test or the MWU test (***p<0.0001, *p<0.05, ns non-significant).

Table 5-9 Descriptive statistic and statistical comparison of the levels of DNA methylation (%) for A) *TNF* and B) *HDAC4* genes in different clinical groups Statistical analyses were performed using the Kruskal-Wallis test followed by Dunn's multiple comparison test.

A TNF

Diagnosis	n	Methylation (%) Median (IQR)
HC	29	6.79 (4.31 - 8.97)
Reactive arthritis	11	7.22 (6.09 - 11.23)
UA	36	6.54 (5.11 - 8.71)
RA	64	3.42 (2.42 - 4.92)
PSA	16	5.49 (8.82 - 6.56)

Comparison test

Kruskal-Wallis chi-squared = 45.286, df = 4, p-value = 3.4×10^{-9}

Dunn (1964) Kruskal-Wallis multiple comparison

Comparison		p value
HC	RA	1.97×10^{-5}
HC	Reactive	1.0000
HC	UA	0.7530
HC	PSA	0.7064
RA	Reactive	4.21×10^{-4}
RA	UA	4.26×10^{-7}
RA	PSA	0.1555

B HDAC4

Diagnosis	n	Methylation (%) Median (IQR)
HC	29	13.83 (10.50 - 21.77)
Reactive arthritis	10	18.63 (8.12 - 42.26)
UA	36	18.90 (12.80 - 22.98)
RA	64	11.73 (9.19 - 16.78)
PSA	16	17.42 (12.10 - 18.50)

Comparison test

Kruskal-Wallis chi-squared = 9.73, df = 4, p-value = 0.0452

Dunn (1964) Kruskal-Wallis multiple comparison

Comparison		p value
HC	RA	1.0000
HC	Reactive	1.0000
HC	UA	1.0000
HC	PSA	1.0000
RA	Reactive	1.0000
RA	UA	0.0257
RA	PSA	1.0000

Modelling of the potential added value of the *TNF* qMSP assay for an early diagnosis of RA

In this section, I determine whether *TNF* methylation detected by our qMSP assay has added value for the early classification of RA using binary logistic regression analysis. In clinical practice, RA patients are classified using criteria developed by the American College of Rheumatology and the European League Against Rheumatism developed (353). In this criteria set, RA is defined based on the presence of synovitis in at least one joint that cannot be explained by the other diagnosis, and achievement a total score of ≥6 out of 10 points. The score comes from 4 domains: number and site of involved joints (range 0–5), serological abnormality (ACPA or RF, range 0–3), elevated acute-phase response (range 0–1), and symptom duration (range 0–1) (353). Using these criteria helped diagnosing RA earlier but there is still a need for more biomarker especially in ACPA negative patients.

To determine whether *TNF* methylation levels have a significant contribution to classification I used a model including other demographic and clinical variables such as age, gender, smoking, autoAb status, TJC, SJC, CRP or DAS. The autoantibodies RF and ACPA which are already part of 2010 RA Classification criteria were excluded from analysis variables to avoid redundancy and over fitting modeling with ACPA notably.

The analysis was performed first in the overall patient group and then separately for ACPA negative patients, which is the group that would have most benefit from my work.

Univariate predictive value of theTNF qMSP assay

The *TNF* qMSP results obtained for 127 DNA samples from patients classified as 64 RA and 63 non-RA (included UA, reactive and PSA) (after 2 poor quality samples were removed as described earlier), were used in this analysis.

The different in *TNF* methylation levels (%) between RA and Non-RA group was highly significant (Δ-DM= -3.02%, p= 4.1×10^{-9}), with lower values in RA group. (Figure 5-28, A). I then performed a predictive analysis choosing a univariate logistic regression approach to establish the characteristics of the qMSP assay alone (unadjusted).

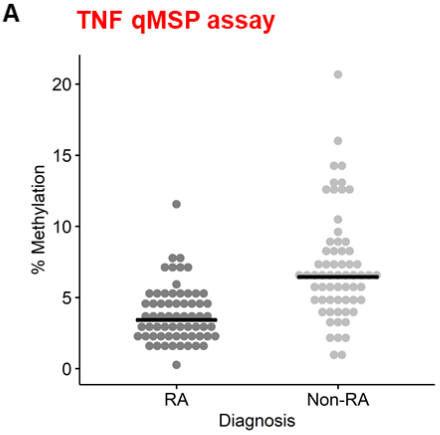
First, the *TNF* methylation level (%) variable was tested in an binary logistic regression to determine its relationship with the diagnosis of RA. Higher levels of methylation being protective, the OR for being RA was 0.62 (95% CI: 0.49 - 0.74), $p = 2.1 \times 10^{-6}$ suggesting the odds of being RA reduces by 0.62 with each unit increase in *TNF* methylation. The ROC (Figure 5-28, B) was built to describe the performance of the assay. The model showed an AUROC =0.197 (95% CI: 0.119 - 0.276) which can be considered a good classification model between the RA and Non-RA groups.

A second analysis was performed using these results after dichotomization to determine if this could improve the performance of the model. This used the results of the *TNF* qMSP assay as a risk category: high and low risk, based on above *TNF* qMSP regression model. Using a cut-off of 80% specificity (a clinically acceptable risk) corresponding to a value of 4.49% of methylation, a new variable for *TNF* methylation level was created (as below / above 4.49% methylation) and used to define the risk categories as high (below) versus low risk (above).

The unadjusted logistic regression and ROC curve analysis were performed using the ***TNF* risk** category (Figure 5-28,C and D). The OR for being RA using the high/low-risk category was 8.46 (95%CI: 3.87 - 19.58, $p = 2.16 \times 10^{-7}$), which was better than using the continuous variable. On the other hand the AUROC of *TNF* risk category (0.741, 95%CI: 0.664, 0.817) was less good than *TNF* qMSP results model, suggesting slightly lower classification performance.

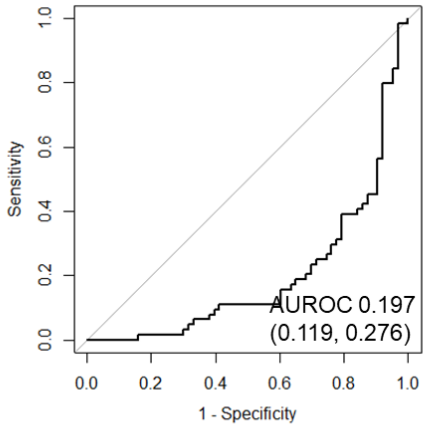
Descriptive performances of the qMSP risk category were calculated (displayed on Figure 5-28,C). Sensitivity was 67.8% suggesting that over than half of the study population shows positivity (high risk) for the biomarker. PPV and NPV were 77.2 and 71.4%, respectively. This mean 77.2% of the individual positive for the marker are actual being RA, and 71.4 % of the negative individual are truly not RA patients. Altogether this suggest good discrimination and 74.0% of individual in the groups being correctly diagnosed based on the risk associated with the patients.

A

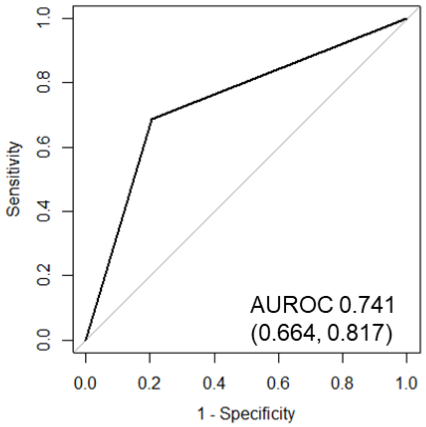


	Medain (IQR)
RA (n=64)	3.41 (2.42, 4.92)
Non-RA (n=63)	6.44 (4.83, 8.39)
MWU test :	
p-value	$4.1x10^{-9}$
Logistic Regression:	
OR	0.62 (0.49, 0.74)
p-value	2.1×10^{-6}

B



D



C

TNF qMSP assay (risk category)

Cut-off value	4.49%
High risk/ Low risk	57 / 70
Logistic Regression:	
OR	8.46 (3.87, 19.58)
p-value	2.16×10^{-7}

Confusion matrix:

	Actual values	
Predicted values	RA	Non-RA
High-risk	44	13
low-risk	20	50
Specificity	79.4 (67.3, 88.5)	
Sensitivity	67.8 (55.9, 79.7)	
PPV	77.2 (64.2, 87.3)	
NPV	71.4 (59.4, 81.6)	
Accuracy	74.0 (65.5,81.4)	

Figure 5-28 Univariate predictive value of the *TNF* qMSP assay

A) qMSP *TNF* methylation data for the RA and Non-RA group. Crossbar presents median of methylation level (%). The statistical analysis compares the difference between groups was performed using the MWU test. The association between *TNF* methylation and the classification of RA was determined using unadjusted logistic regression determining an odd ratio (OR)

B) The overall performance of the *TNF* qMSP levels was determined by an AUCROC (95% CI) analysis.

C) *TNF* methylation risk category defined as high/row risk of RA using a cutoff of the continuous levels of methylation at 4.49% for 80% specificity. The confusion matrix of the classification results compared to the actual diagnosis result is displayed. Sensitivity, specificity, NPV, and PPV are described in the table. The relationship between the *TNF* risk categories and RA was determined using unadjusted logistic regression.

D) The overall performance of the *TNF* qMSP risk categorization was determine by an AUCROC (95% CI) analysis.

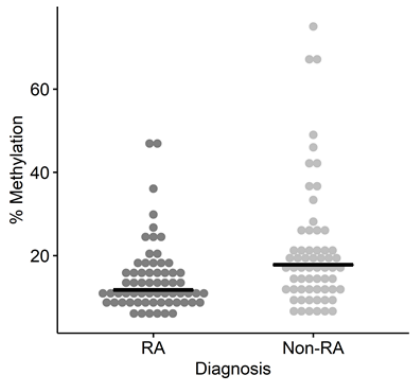
Univariate value of the HDAC4 qMSP data

A similar analysis was performed for *HADAC4* qMSP data. The difference in *HDAC4* methylation levels (%) between RA and Non-RA group was significant (Δ-DM = -6.05%, MWU p = 0.0016), with lower values in RA group. (Figure 5-29, A). An unadjusted binary logistic regression analysis gave the OR of 0.95 (95% CI: 0.91 - 0.98, p = 0.0064) suggesting that *HDAC4* methylation has a small value for the classification of RA. The AUROC =0.336 (95% CI: 0.241 - 0.432, Figure 5-29, B) was not particularly good and furthermore, close to 0.5 (random classifier).

The second analysis using *HDAC4* risk category defined a cut-off (80% specificity) corresponding to a methylation value of 10.66 %, was performed. The unadjusted logistic regression gave a non-statistical significant OR value for being RA of 2.67 (95%CI: 1.21 - 6.14, p =0.0168, Figure 5-29, C) and the AUROC of 0.599 (95%CI:0.521 - 0.677, Figure 5-29, D) suggesting that *HDAC4* high/low-risk category had less value for the classification RA and non-RA groups.

Compared to the *TNF* qMSP assay, the *HDAC4* assay showed a very limited capacity to distinguish between RA and non-RA groups. Therefore no further analysis of both the continuous and categorical variables resulting from the *HDAC4* assay were performed.

A **HDAC4 qMSP assay**



	Medain (95%CI)
RA (n=64)	11.73 (9.19, 16.78)
Non-RA (n=62)	17.78 (11.01, 21.82)

MWU test :

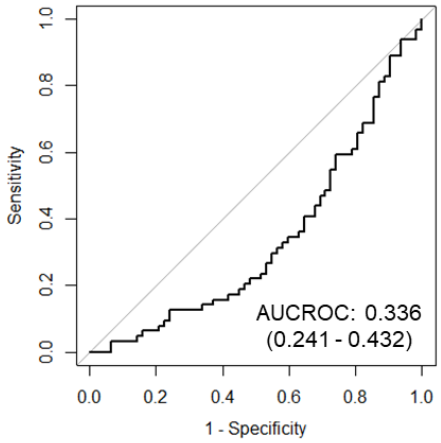
p-value 0.0016

Logistic Regression:

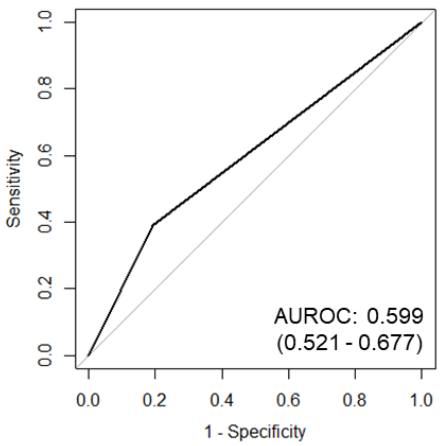
OR 0.95 (0.91, 0.98)

p-value 0.0064

B **HDAC4 qMSP assay ROC**



D **Risk category ROC**



C

HDAC4 qMSP assay (risk category)

Cut-off value 10.66%

High risk/ Low risk 37 / 89

Logistic Regression:

OR 2.67 (1.21, 6.14)

p-value 0.0168

Confusion matrix:

	Actual values	
Predicted values	RA	Non-RA
High-risk	25	12
low-risk	39	50
Specificity	80.6 (68.6, 89.6)	
Sensitivity	39.1 (21.1, 52.1)	
PPV	67.6 (50.2, 81.9)	
NPV	56.1 (45.3, 66.7)	
Accuracy	59.5 (50.4, 68.2)	

Figure 5-29 Univariate predictive value of the *HDAC4* qMSP assay

A) qMSP *HDAC4* methylation data for the RA and Non-RA group. Crossbar presents median of methylation level (%). The statistical analysis compares the difference between groups was performed using the MWU

test. The association between *HDAC4* methylation and the classification of RA was determined using unadjusted logistic regression determining an odd ratio (OR)

B) The overall performance of the *HDAC4* qMSP levels was determined by an AUCROC (95% CI) analysis.

C) *HDAC4* methylation risk category defined as high/row risk of RA using a cutoff of the continuous levels of methylation at 10.66% for 80% specificity. The confusion matrix of the classification results compared to the actual diagnosis result is displayed. Sensitivity, specificity, NPV, and PPV are described in the table. The relationship between the *HDAC4* risk categories and RA was determined using unadjusted logistic regression.

D) The overall performance of the *HDAC4* qMSP risk categorization was determine by an AUCROC (95% CI) analysis.

Adjusted analysis of the *TNF* qMSP assay with other parameters

Altogether, the unadjusted analysis suggested that the *TNF* DNA methylation risk category was better for the prediction based on OR but less so based on AUROC results, therefore further analysis will keep testing both the continuous levels and the risk category variables to establish the added value of using a *TNF* qMSP assay. Several demographic and clinical parameters are known to be associated with RA classification (2, 385-387) and could present co-variance with the qMSP assay, therefore limiting its added value while individually, it is highly predictive.

Demographic and clinical parameters data recorded in this group of 127 patients include age, gender, smoking, positivity for RF, ACPA, tender joints count (TJC), swollen joints count (SJC), CRP level, a routinely used disease activity score (DAS28), and symptoms duration at inclusion for each patients. These were retrieved from the IACON and EAC database.

The descriptive and the individual statistical significance of these data in a comparison between RA and non-RA group (MWU and X^2 tests) are presented in Table 5-10 and plotted in Figure 5-30. All parameters apart from gender and symptom duration were significantly different between the RA and Non-RA group. Age and smoking are known risk factor associate with RA, as well as the disease characteristic variables (TJC, SJC, CRP), and DAS28, which were higher in RA group. The symptom duration which is also part of the chronic inflammatory disease classification, show no difference between RA and non-RA groups as these were all early IA patients awaiting a classification. The autoantibodies, RF and ACPA which are the serological biomarker used as classification criteria are obviously associated with RA patients. Because of this, autoantibodies were excluded from further multivariate analysis. Altogether, this small groups of 127 patients was representative of the wider population of people with early IA symptoms.

Table 5-10 Description of demographic and clinical parameters in overall patient (both APCA positive and negative groups). The descriptive statistic (median (IQR) or frequency) and statistical significant (MWU or X^2 tests) comparison between RA and non-RA group are presented. * indicated variables with a few missing data.

Variable	Non-RA n= 63	RA n= 64	p-value
Age	40 (33-49)	61.5 (46.75-73)	1.91×10^{-9}
Gender (M/F)	18/45	13/51	0.3806
Smoking (never/smoker)	32/25 *	20/40 *	0.0217
RF (positive/negative)	6/55 *	31/33	5.90×10^{-6}
ACPA (positive/negative)	2/61	29/31 *	2.72×10^{-8}
Duration	6 (3-12)	4 (2-8)	0.2595
Tender joint count	2 (1-6)	9 (5-15.25)	2.64×10^{-7}
Swollen joint count	1 (0-2)	5.5 (1.75-9	3.91×10^{-7}
CRP	0 (0-10.8)	12 (0-23.25)	0.0022
DAS28	2.9 (2.3-3.95)	5.05 (3.68-5.63)	9.6×10^{-8}

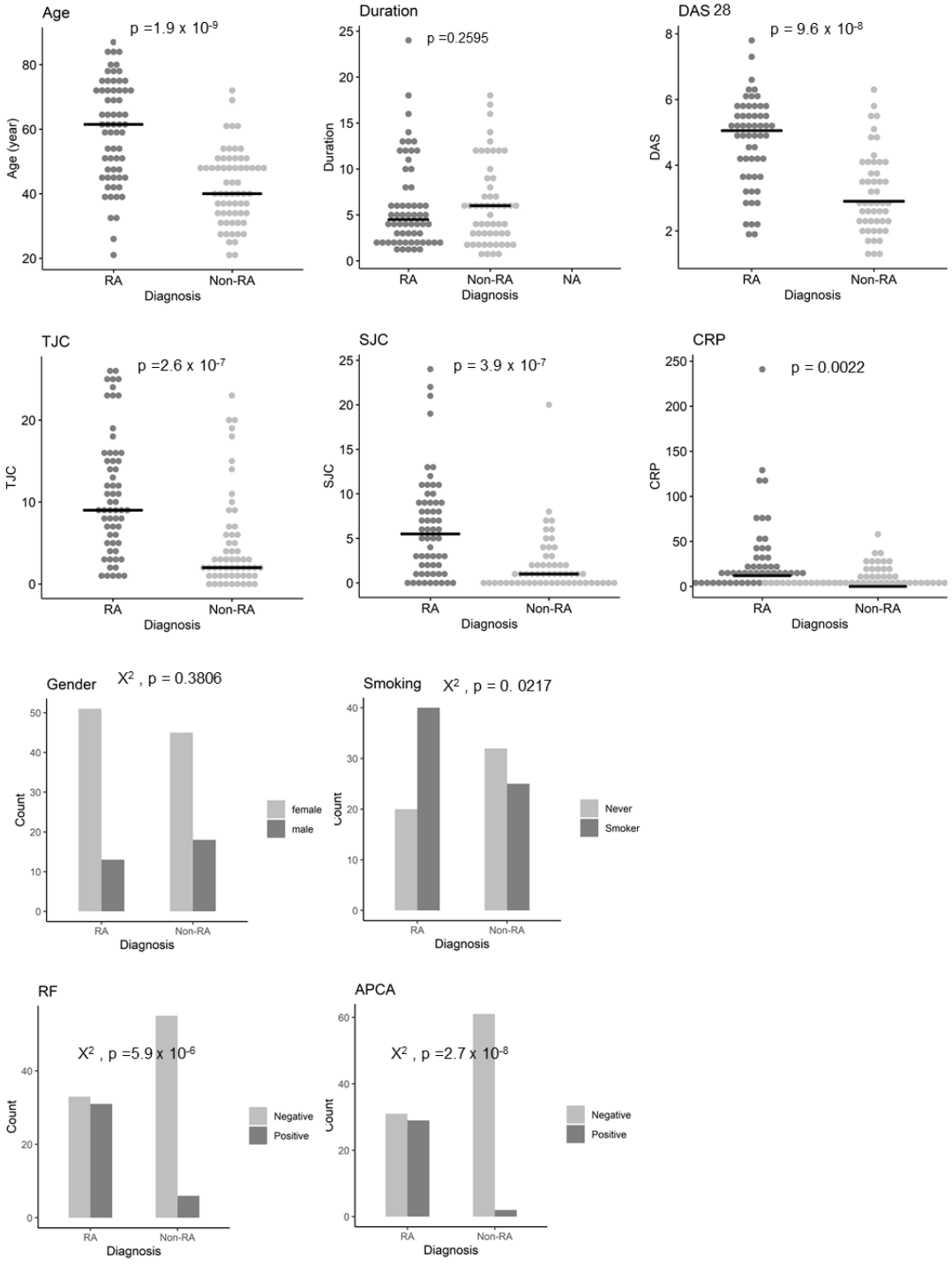


Figure 5-30 Demographic and clinical parameters of 127 patients shown as dot plot (for continuous variables) and the frequency bar plot (for categorical variables). Statistical significance (MWU or X^2 tests) comparison RA and non-RA group are presented.

Binary logistic regressions were first performed for all variables individually, to determine their unadjusted relationship with the diagnosis of being RA. Unadjusted p-value and OR of individual factor are listed in 1st column of Table 5-11.

The variables that affected RA (OR not equal to 1) and statistically significant (p-value <0.001) included age, smoking, RF, ACPA, TJC, SJC, CPR, and DAS28 while gender and symptom duration were not. Of note, in this small cohort, the highest significant OR were obtained for APCA and RF as expected.

ROC curve for these variables were generated (Figure 5-31) and the descriptive characteristics of the individual variable model using a cut-off at 80% specificity were calculated (specificity, PPV, NPV) (Table 5-11). The prediction value of some clinical variable such as RF, ACPA, and CRP was particularly high (AUROC 0.693, 0.726, and 0.654, respectively) in this small group of 127 patients.

RF and ACPA which are specific biomarkers for RA showed the strongest association with RA as expected. The other known risk factor (i.e. age, gender, and smoking) and the disease activity parameters also showed association with RA.

I then verified whether there was any association between the levels of *TNF* gene DNA methylation (%) and demographic or clinical variables to address possible cofounding effects. I use Point biserial correlation for categorical variables (gender, RF, ACPA, smoking) and Spearman correlation test for continuous variables (age, symptom duration, TJC, SJC, CRP and DAS28). The correlation coefficient between *TNF* qMSP result and all demographic or clinical variables are low and not statistically significant (Table 5-12). Although age showed a correlation coefficient (rho= -0.41, p= 2.01 x 10^{-6}), it was still below the 0.600 threshold for being considered a relevant relationship with *TNF* methylation.

This result suggested that there was no obvious relationship between *TNF* methylation and other parameters. The association between *TNF* methylation and RA status may therefore be not redundant with any other risk factors, thus allowing to have added value in a model already using these variables.

Table 5-11 Logistic regression of individual clinical parameters in 125 patients. Regression details, overall performance, and performance using categorization (at a cut-off at 80% specificity) are illustrated.

Independent Variable	Regression OR (95% CI)	p-value	AUROC (95% CI)	Cut-off value	Sensitivity (95% CI)	Specificity (95% CI)	PPV (95% CI)	NPV (95% CI)
Age	1.09 (1.06, 1.13)	1.07×10^{-7}	0.809 (0.733, 0.884)	51.0	68.8 (55.9, 79.7)	80.9 (69.1, 89.8)	78.6 (65.9, 86.8)	71.8 (59.5, 84.0)
Smoking: smoker	2.56 (1.22, 5.48)	0.0140	0.614 (0.525, 0.703)	smoker	66.7 (53.3, 78.3)	56.1 (42.4, 69.3)	61.5 (47.8, 74.3)	61.5 (47.7, 73.8)
RF : positive	8.61 (3.44, 24.89)	1.5×10^{-5}	0.693 (0.621, 0.765)	positive	48.4 (35.8, 61.3)	90.2 (79.8, 96.3)	83.8 (69.0, 89.7)	62.5 (49.7, 82.6)
ACPA: positive	28.53 (7.89, 183.84)	1.14×10^{-5}	0.726 (0.658, 0.793)	positive	48.3 (35.2, 61.6)	96.8 (88.9, 99.6)	93.5 (79.4, 96.1)	66.3 (53.3, 94.3)
TJC	1.15 (1.08, 1.24)	3.55×10^{-5}	0.775 (0.689, 0.861)	8.0	65 (51.6, 76.9)	80.7 (68, 89.9)	78 (64.4, 86.4)	68.6 (55.7, 82.4)
SJC	1.30 (1.16, 1.49)	4.14×10^{-5}	0.775 (0.689, 0.861)	4.0	58.3 (44.9, 70.9)	80.7 (68.1, 89.9)	76.1 (61.9, 84.7)	64.8 (51.7, 79.8)
CRP	1.03 (1.01, 1.06)	0.0079	0.654 (0.561, 0.748)	18.0	37.5 (25.7, 50.4)	80.0 (67.7, 89.2)	66.7 (51.1, 77.3)	54.5 (40.9, 71.3)
DAS28	2.37 (1.71, 3.46)	1.24×10^{-6}	0.801 (0.716, 0.886)	4.2	71.4 (57.8, 82.7)	82.0 (68.5, 91.4)	81.6 (68.0, 89.5)	71.9 (58.4, 85.7)
Gender: male	0.63 (0.27, 1.43)	0.2800	NA					
Symtom duration	0.98 (0.94, 1.01)	0.3020	NA					

Table 5-12 The relationship between TNF methylation and demographic or clinical variables. The distributions of the variables were tested for normality using the Shapiro-Wilkinson test. For the continuous variables, the correlation was tested using Spearman's correlation due to the skewed distribution. For the categorical variable, the correlation was tested using point biserial correlation.

Variable	Coefficience	p-value
Age	-0.41	0.0000
Smoking	-0.20	0.0319
RF	-0.21	0.0201
ACPA	-0.16	0.0846
TJC	-0.28	0.0019
SJC	-0.32	0.0005
CRP	-0.15	0.0972
DAS28	-0.33	0.0006
Gender	-0.09	0.3284
Symtom duration	0.12	0.1764

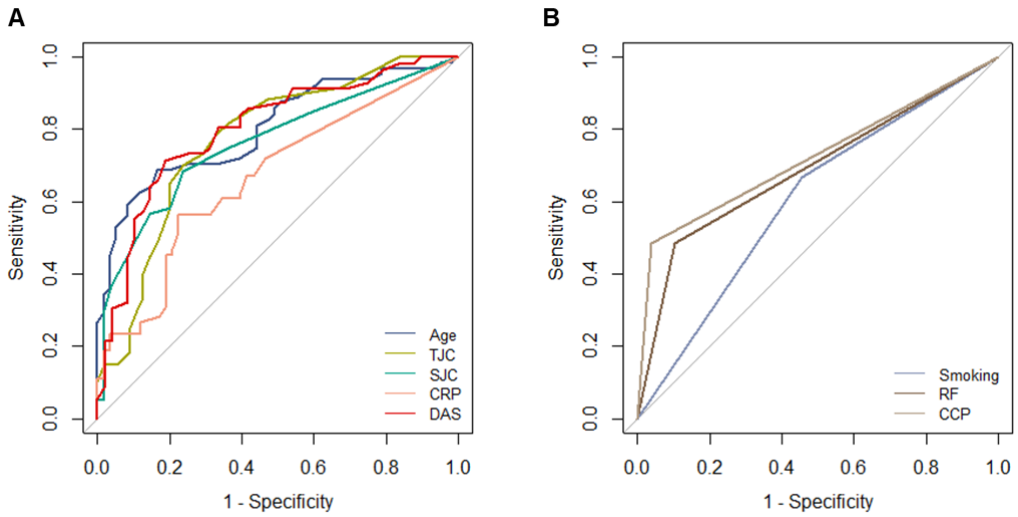


Figure 5-31 ROC curve of demographic and clinical parameters of overall patient for ; (A) continuous variables (B) categorical variables.

The individual characteristics of the *TNF* qMSP assay (Figure 5-28) suggest that they have similar performance to demographic and clinical variables and may therefore contribute to improving models. All variable that showed an association with RA in unadjusted logistic regression (excluding RF and ACPA which are part of the 2010 EULAR RA classification criteria) were used in a multiple logistic regression allowing several variables to be considered at the same time.

Different models using multiple binary logistic regression including only the clinical variables to be considered were first developed. Because DAS28 utilize CRP and TJC / SJC, It would be redundant if DAS28 and those were included in the same model. Thus, two types of model were developed, one including the significant variables age and smoking with TJC, SJC, and CPR (clinical model 1), and another model using age and smoking with DAS28 instead (clinical model 2). All variable used in each model and the model performance are described in Table 5-13.

OR in the clinical model 1 showed the dominant contribution of Age (OR =1.07, p=0.0003) while clinical model 2 showed both age (OR = 1.07, p=0.0003) and DAS28 (OR = 1.91, p=0.0008). Clinical Model 2 also gave a slightly better fit (higher R^2) and shows slightly better performance for discriminating RA and Non-RA (higher AUROC). This implies that using DAS28 instead of TJC, SJC, and CPR might be a better option in predicting RA in this particular group of patients as it is relatively small (n=127). However, this effect was not clear thus I kept testing both models in further analysis.

I developed a second version of each model (model-1.2 and model-2.2) adding the *TNF* qMSP assay results into the clinical models to see whether it improved the overall model performance. Adding *TNF* to the clinical model 1 and 2 improved the model's performance (higher AUROC and R^2). AUROC increase from 0.856 to 0.902 in clinical model 1 and from 0.861 to 0.904 in clinical model 2.

For a third version of each model (model-1.3, model-2.3), I added the *TNF* variable as categorical data (*TNF* risk category) to see whether this improved the model performance compared to the continuous variable. Adding *TNF* risk category to the clinical model 1 and 2 improved the model's performance of clinical model alone (AUROC are 0.881 and 0.895 for the clinical model 1.3 and 2.3, respectively). This suggested that using the categorical *TNF* measurement (risk category) had added value for both clinical models, however, the continuous *TNF* methylation level (*TNF* qMSP model-1.2 and model-2.2) showed better performance in classifying RA and non-RA groups.

In a last version of each model (model-1.4, model-2.4) I tested the impact of adding the *HDAC4* qMSP data on top of the *TNF* qMSP data. *HDAC4* was not providing any added value to model 1 but maybe a very small (neglectable) improvement in model 2 (AUROC +0.001).

Table 5-13 Different multiple logistic regression models of overall patient show the add up value of qMSP assays to the clinical model.

Overall patient

	Clinical model 1				Clinical model 1 + TNF qMSP				Clinical model 1 +TNF +HDAC4 qMSP			
	Model 1				**Model 1.2**				**Model 1.3**			
Variable	OR	(95% CI) 2.5 %	97.5%	p value	OR	(95% CI) 2.5 %	97.5%	p value	OR	(95% CI) 2.5 %	97.5%	p value
Age	1.07	1.03	1.11	0.0003	1.06	1.02	1.11	0.0017	1.06	1.03	1.12	0.0017
Smoking: smoker	1.54	0.57	4.17	0.3899	1.42	0.49	4.16	0.5172	1.42	0.48	4.16	0.5209
TJC	1.08	1.00	1.18	0.0516	1.08	0.99	1.19	0.0782	1.08	0.99	1.19	0.0772
SJC	1.08	0.94	1.27	0.2808	1.05	0.90	1.23	0.5626	1.05	0.91	1.24	0.5415
CRP	1.02	1.00	1.06	0.1702	1.02	0.99	1.06	0.2248	1.02	0.99	1.06	0.2111
TNF qMSP					0.69	0.55	0.85	0.0012	0.68	0.52	0.85	0.0018
HDAC4 qMSP									1.01	0.95	1.08	0.6519
	R^2	0.492			R^2	0.589			R^2	0.591		
	AUROC	**0.856**			AUROC	**0.902**			AUROC	**0.902**		

	Clinical model 1 + TNF risk category			
	Model 1.3			
Variable	OR	(95% CI) 2.5 %	97.5%	p value
Age	1.07	1.03	1.11	0.0012
Smoking: smoker	1.17	0.40	3.36	0.7738
TJC	1.09	1.00	1.19	0.0565
SJC	1.06	0.92	1.25	0.4261
CRP	1.02	0.99	1.05	0.3017
TNF risk category	4.77	1.75	13.75	0.0028
	R^2	0.558		
	AUROC	**0.881**		

	Clinical model 2				Clinical model 2 + TNF qMSP				Clinical model 2 +TNF +HDAC4 qMSP			
	Model 2				**Model 2.2**				**Model 2.2**			
Variable	OR	(95% CI) 2.5 %	97.5%	p value	OR	(95% CI) 2.5 %	97.5%	p value	OR	(95% CI) 2.5 %	97.5%	p value
Age	1.07	1.03	1.11	0.0003	1.07	1.03	1.12	0.0025	1.07	1.03	1.12	0.0024
Smoking: smoker	1.26	0.44	3.51	0.6598	1.12	0.36	3.44	0.8376	1.12	0.35	3.44	0.8442
DAS28	1.91	1.33	2.85	0.0008	1.85	1.24	2.89	0.0037	1.87	1.25	2.93	0.0035
TNF qMSP					0.67	0.52	0.83	0.0009	0.66	0.50	0.83	0.0013
HDAC4 qMSP									1.02	0.95	1.08	0.6294
	R^2	0.497			R^2	0.610			R^2	0.611		
	AUROC	**0.861**			AUROC	**0.904**			AUROC	**0.905**		

	Clinical model 2 + TNF risk category			
	Model 2.3			
Variable	OR	(95% CI) 2.5 %	97.5%	p value
Age	1.07	1.03	1.12	0.0010
Smoking: smoker	0.92	0.29	2.77	0.8786
DAS28	1.82	1.22	2.83	0.0047
TNF risk category	5.43	1.89	16.82	0.0022
	R^2	0.572		
	AUROC	**0.895**		

Overall the best model (2.2) was that including DAS28 and the *TNF* qMSP data (AUROC 0.904) probably as fitting best the small number of patients (n=127) compared to the model 1.2. However in all models, some variables contributed non-significantly (smoking, CRP, TJC and SJC in model 1 and smoking in model 2). I re-run these models using an forward regression approach which suggested to keep the same variables indeed but improved slightly the OR and p-values of the remaining variables but showed a the slightly less good AUROC.

Model 1

	Age				**Age + TNF qMSP**				**Age +TNF risk**			
Variable	OR	(95% CI) 2.5 %	97.5%	p value	OR	(95% CI) 2.5 %	97.5%	p value	OR	(95% CI) 2.5 %	97.5%	p value
Age	1.09	1.06	1.34	1.07×10^{-7}	1.08	1.05	1.13	1.75×10^{-5}	1.09	1.05	1.13	4.1×10^{-6}
TNF					0.66	0.53	0.81	1.58×10^{-4}	6.39	2.61	16.56	7.27×10^{-5}
	R^2	0.386			R^2	0.527			R^2	0.505		
	AUROC	**0.809**			AUROC	**0.876**			AUROC	**0.869**		

Model 2

	Age+DAS28				**Age + DAS28 + TNF qMSP**				**Age + DAS28+ TNF risk**			
Variable	OR	(95% CI) 2.5 %	97.5%	p value	OR	(95% CI) 2.5 %	97.5%	p value	OR	(95% CI) 2.5 %	97.5%	p value
Age	1.07	1.04	1.12	0.0003	1.07	1.03	1.12	0.0023	1.07	1.03	1.12	0.0010
DAS28	1.94	1.36	2.88	0.0005	1.87	1.26	2.90	0.0029	1.81	1.22	2.78	0.0042
TNF					0.67	0.52	0.83	0.0089	5.34	1.89	16.17	0.0020
	R^2	0.495			R^2	0.610			R^2	0.572		
	AUROC	**0.863**			AUROC	**0.903**			AUROC	**0.894**		

Analysis repeated in ACPA negative patient

As mention earlier the ACPA test, used for diagnosis of RA nowadays, (and RF to a lower extend) is the most useful biomarker. People who test positive for ACPA are very likely to develop RA (98% specificity (356)), but only about half of people with RA show positivity for this this antibody (~50-60% sensitivity in population cohorts (356, 358, 359)). This result in ACPA-negative patients experiencing delays in getting an early diagnosis and more importantly an early treatment. A novel diagnosis biomarker with an ability to classify patients with RA and Non-RA, especially in ACPA-negative patient group is of great clinical importance.

The *TNF* qMSP biomarker evaluation was performed again in ACPA negative patients (n=92) of which 31 were classified as RA (which was indeed about half of the overall RA patients) and 61 were Non-RA. The same analysis process was repeated in this group. The description and the individual statistical significance of the demographic and clinical data in a comparison between RA and non-RA group in APCA negative group are presented in Table 5-14 (MWU and X^2 tests) and Figure 5-32.

The *TNF* methylation was different in RA and significantly lower than Non-RA group (Δ-DM = -3.28%, $p=1.4x10^{-8}$) (Figure 5-33). Unadjusted logistic regression analysis was performed. *TNF* qMSP still predicted RA in ACPA negative group with an OR 0.49 (95%CI: 0.34 - 0.65, $p =1.25x10^{-5}$) and an AUROC of 0.136 (0.059-0.214). Similar to the overall patient group, *TNF* risk category had an OR of 14 (5.12- 42.8) and an AUROC = 0.789 (0.699-0.879) did not help improve model ability compared to *TNF* qMSP.

Table 5-14 Description of demographic and clinical parameters in ACPA negative patient group. The descriptive statistic (median (IQR) or frequency) and statistical significant (MWU or X^2 tests) comparison between RA and non-RA group are presented. * indicated variables with a few missing data.

Variable	Non-RA n= 61	RA n= 31	p-value
Age	41 (34-50)	65 (53.5-75)	1.66×10^{-9}
Gender (M/F)	18/43	4/27	0.1319
RF (positive/negative)	5/55 *	8/23	0.0522
Smoking (never/smoker)	31/24 *	12/19	0.1778
Duration	6 (3-12)	4 (2-6)	0.0225
Tender joint count	3 (1-6)	11 (7-15.5)	7.43×10^{-6}
Swollen joint count	1 (0-2)	7 (2.5-11)	1.81×10^{-6}
CRP	2.5 (0-11.6)	8 (0-19.2)	0.0832
DAS28	2.9 (2.3-4)	5.1 (4.15-5.9)	1.27×10^{-5}

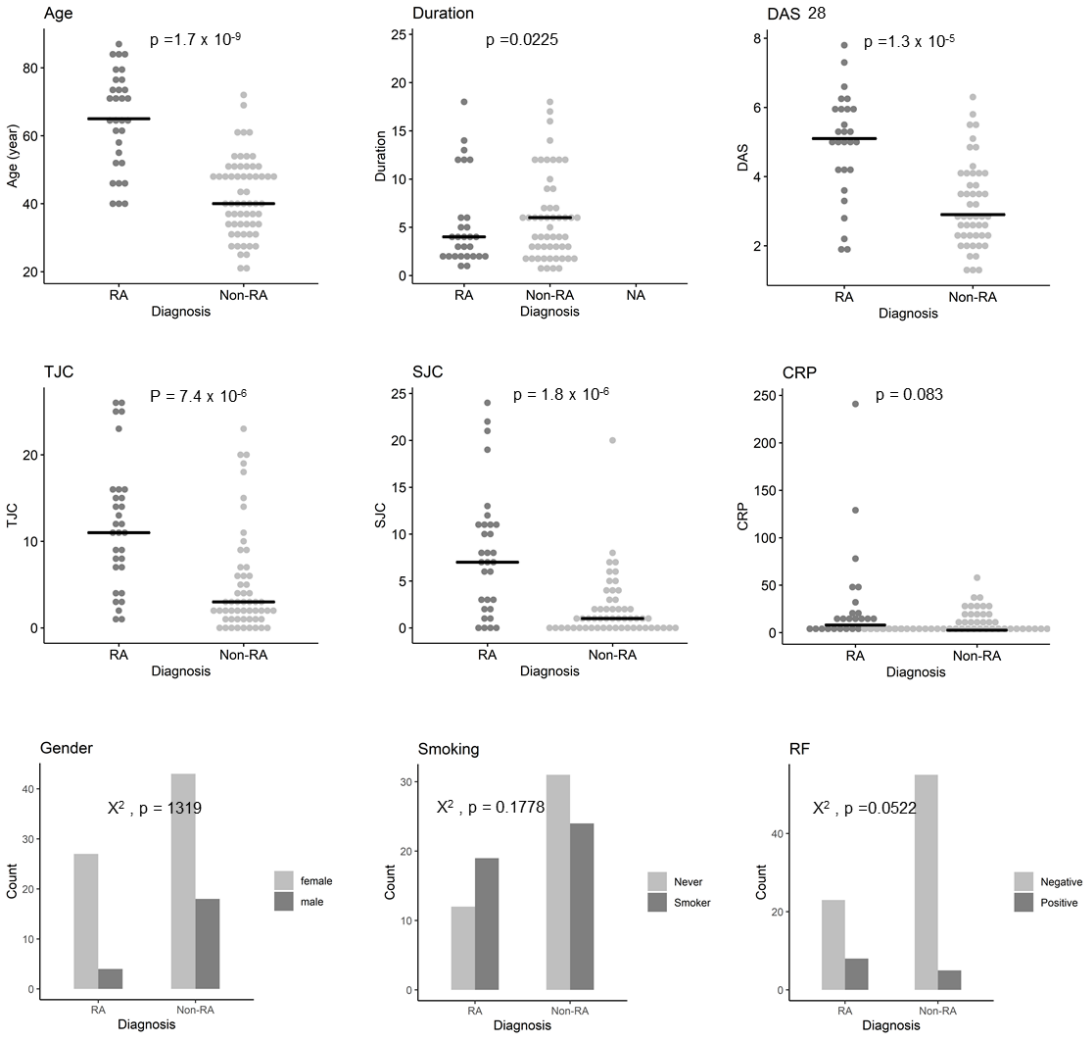
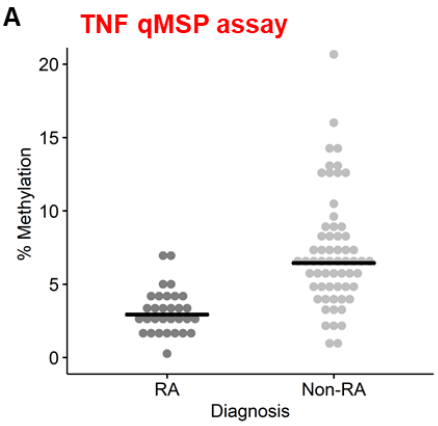
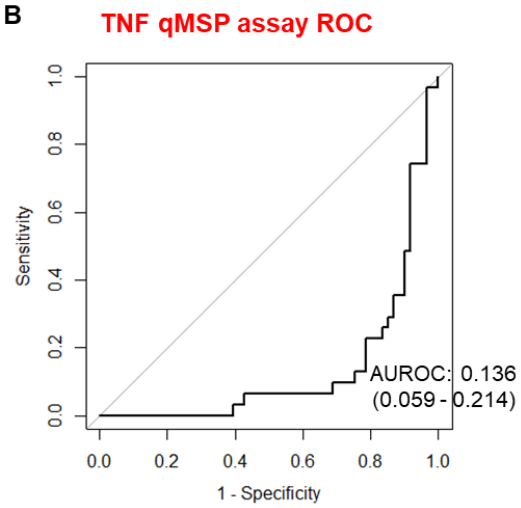
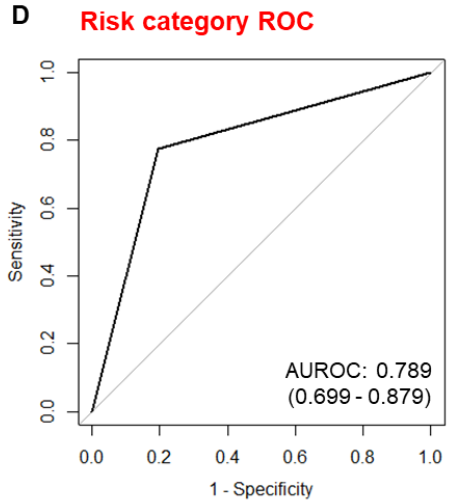


Figure 5-32 Demographic and clinical parameters of ACPA negative patients shown as dot plot (for continuous variables) and the frequency bar plot (for categorical variables). Statistical significance (MWU or X^2 tests) comparison RA and non-RA group are presented.



	Medain (95%CI)
RA (n=31)	3.16 (2.14, 3.92)
Non-RA (n=61)	6.44 (4.73 - 8.55)
MWU test :	
p-value	1.4×10^{-8}
Logistic Regression:	
OR	0.49 (0.34, 0.65)
p-value	1.29×10^{-5}





C

TNF qMSP assay (risk category)

Cut-off value	4.16%
High risk/ Low risk	36 / 56
Logistic Regression:	
OR	14 (5.12, 42.8)
p-value	8.8 x 10-7

Confusion matrix:

	Actual values	
Predicted values	RA	Non-RA
High-risk	24	12
low-risk	7	49
Specificity	80.3 (68.2, 89.4)	
Sensitivity	77.4 (58.9, 90.4)	
PPV	66.7 (49.0, 81.4)	
NPV	87.5 (75.9, 94.8)	
Accuracy	79.3 (69.6, 87.1)	

Figure 5-33 Univariate predictive value of the *TNF* qMSP assay for ACPA negative patients

A) qMSP *TNF* methylation data for the RA and Non-RA group. Crossbar presents median of methylation level (%). The statistical analysis compares the difference between groups was performed using the MWU

test. The association between *TNF* methylation and the classification of RA was determined using unadjusted logistic regression determining an odd ratio (OR)

B) The overall performance of the *TNF* qMSP levels was determined by an AUCROC (95% CI) analysis.

C) *TNF* methylation risk category defined as high/row risk of RA using a cutoff of the continuous levels of methylation at 4.16% for 80% specificity. The confusion matrix of the classification results compared to the actual diagnosis result is displayed. Sensitivity, specificity, NPV, and PPV are described in the table. The relationship between the TNF risk categories and RA was determined using unadjusted logistic regression.

D) The overall performance of the *TNF* qMSP risk categorization was determine by an AUCROC (95% CI) analysis.

Similarly, unadjusted binary logistic regression was performed for all other clinical variables (Table 5-15 and Figure 5-34). The variables that predicted RA include age, TJC, SJC, CRP, DAS28, TNF methylation. RF and smoking variables were no longer significantly associated with RA (OR p-value > 0.05) in this ACPA negative group.

Two multiple logistic regression models were developed using age and either TJC, SJC, CRP (clinical model 3) or DAS28 (clinical model 4). Similar to the overall patient group, age still be the variable that associated with RA the most.

TNF qMSP was added in both model (model-3.2 and model-4.2) (Table 5-16) and resulted in an AUROC improvement of +4.2% for model 3 and of +4.7 % for model 4.

TNF risk category was also added in in both model (model-3.3 and model-4.3) (Table 5-16). Compared to *TNF* qMSP, the *TNF* risk category showed less improvement to both clinical model (AUROC increased by +3.3% for model 3 and of +3.9 % for model 4).

Table 5-15 Logistic regression of individual clinical parameters in ACPA negative patients. Regression details, overall performance, and performance using categorization (at a cut-off at 80% specificity) are illustrated.

	Regression				Sensitivity	Specificity	PPV	NPV
Independent Variable	**OR (95% CI)**	**p-value**	**AUROC (95% CI)**	**Cut-off value**	**(95% CI)**	**(95% CI)**	**(95% CI)**	**(95% CI)**
Age	1.14 (1.09, 1.21)	9.07×10^{-7}	0.886 (0.814, 0.958)	51.0	80.1 (62.5, 92.5)	80.3 (68.1, 89.4)	67.6 (52.2, 86.1)	89.1 (76.6, 94.4)
Smoking: smoker	2.04 (0.84, 5.12)	0.1184	NA	smoker				
RF : positive	3.82 (1.15, 13.86)	0.0309	0.587 (0.502, 0.673)	positive	25.8 (11.9, 44.6)	91.7 (81.6, 97.3)	61.5 (39.2, 78.7)	70.5 (48.0, 88.4)
TJC	1.16 (1.08, 1.25)	0.0001	0.791 (0.693, 0.89)	8.0	71.0 (52.0, 85.8)	80.0 (67.0, 89.6)	66.7 (50.4, 83.1)	83.0 (68.4, 91.4)
SJC	1.32 (1.16, 1.54)	9.37×10^{-5}	0.806 (0.701, 0.911)	4.0	64.5 (45.4, 80.8)	80.0 (67.0, 89.6)	64.5 (48.0, 80.8)	80.0 (64.6, 89.6)
CRP	1.03 (1.00, 1.05)	0.0684	0.607 (0.486, 0.728)	18.4	29.0 (14.2, 48.0)	79.3 (66.6, 88.8)	42.9 (28.1, 62.9)	67.6 (45.9, 81.3)
DAS28	2.39 (1.63, 3.79)	4.42×10^{-5}	0.804 (0.691, 0.918)	4.2	74.1 (53.7, 88.9)	81.6 (68.0, 91.2)	69 (51.4, 86.1)	85.1 (70.0, 93.0)
Gender: male	0.35 (0.09, 1.07)	0.0859	NA					
Symtom duration	0.99 (0.95, 1.02)	0.5744	NA					

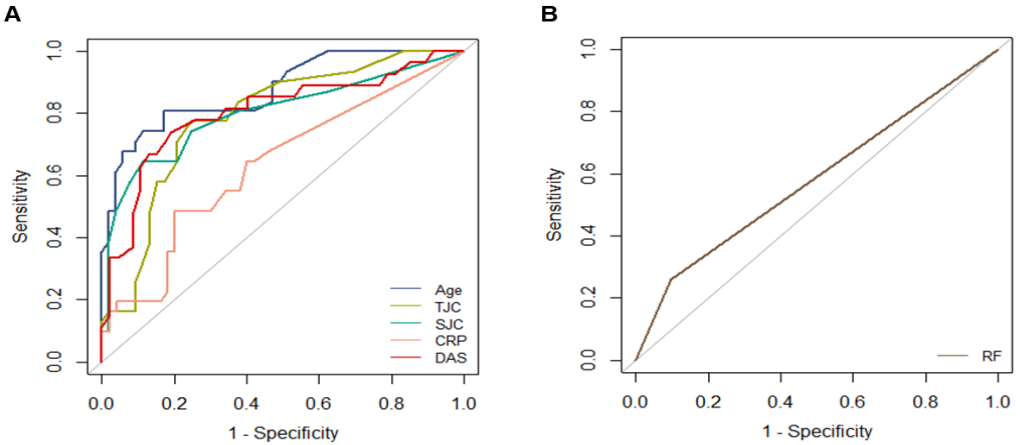


Figure 5-34 ROC curve of demographic and clinical parameters of ACPA negative patient ; A) continuous variables B) categorical variables.

Table 5-16 Different multiple logistic regression models of ACPA negative patients show the add up value of qMSP assays to the clinical model.

APCA negative patient

Clinical model 3

	Model 3			
Variable	OR	(95% CI)		p value
		2.5 %	97.5%	
Age	1.11	1.06	1.18	5.55×10^{-5}
TJC	1.06	0.92	1.22	0.3770
SJC	1.09	0.92	1.32	0.3330
CRP	1.01	0.98	1.06	0.7750
TNF qMSP				
	R^2	0.625		
	AUROC	**0.907**		

Clinical model 3 + TNF qMSP

	Model 3.2			
Variable	OR	(95% CI)		p value
		2.5 %	97.5%	
Age	1.11	1.05	1.19	0.0016
TJC	10.90	0.92	1.28	0.3148
SJC	1.05	0.87	1.31	0.6069
CRP	1.00	0.97	1.04	0.8307
TNF qMSP	0.54	0.34	0.78	0.0031
	R^2	0.735		
	AUROC	**0.949**		

Clinical model 3 + TNF risk category

	Model 3.3			
Variable	OR	(95% CI)		p value
		2.5 %	97.5%	
Age	1.11	1.06	1.19	0.0004
TJC	1.07	0.92	1.24	0.5128
SJC	1.08	0.89	1.32	0.3975
CRP	1.01	0.98	1.06	0.8712
TNF risk category	2.13	0.76	2.78	0.0031
	R^2	0.712		
	AUROC	**0.940**		

Clinical model 4

	Model 4			
Variable	OR	(95% CI)		p value
		2.5 %	97.5%	
Age	1.13	1.07	1.21	0.0001
DAS28	1.60	0.99	2.70	0.0630
TNF qMSP				
	R^2	0.648		
	AUROC	**0.914**		

Clinical model 4 + TNF qMSP

	Model 4.2			
Variable	OR	(95% CI)		p value
		2.5 %	97.5%	
Age	1.12	1.05	1.22	0.0017
DAS28	1.50	0.88	2.71	0.1498
TNF qMSP	0.54	0.32	0.79	0.0053
	R^2	0.757		
	AUROC	**0.961**		

Clinical model 4 + TNF risk category

	Model 4.3			
Variable	OR	(95% CI)		p value
		2.5 %	97.5%	
Age	1.15	1.08	1.27	0.0006
DAS28	1.36	0.75	2.55	0.3153
TNF qMSP	18.13	3.35	162.11	0.0024
	R^2	0.758		
	AUROC	**0.953**		

For both the overall patient and the ACPA negative patient group, comparing the model performances generated from different models suggested that using DAS28 instead of the 3 variable (TJC, SJC, and CRP) allowed for a better prediction, probably due to the fact that we only had 127 patients. The most predictive clinical models without *TNF* (model-2 and model-4) or with TNF (model 2.2 and model 4.2) were directly compared in Figure 5-35. Higher AUROC and R^2 obtained for the model with *TNF* suggests an increase of the overall model performances for the classification of RA. A confusion matrix which compared the model prediction result vs the actual result at a particular probability cut-off, here chosen again at 80% specificity (as a clinically acceptable risk), are showed in the table in Figure 5-35. Descriptive characteristics (specificity, PPV, NPV and accuracy) were calculated. Adding *TNF* qMSP to the clinical model 2 and 4 improve the model accuracy by +2.8 % (from 80.2 to 83.0%) for the overall patient group, and +5.2% (from 80.3 to 85.5%) for the ACPA negative group.

To further confirm the add-up value of the *TNF* methylation assay to the clinical model, an analysis was performed to test the difference between the AUC of the 2 models, with and without *TNF*.

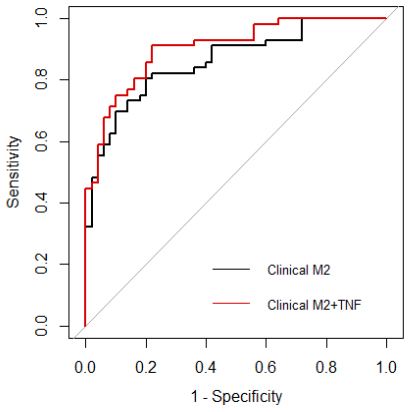
Bootstrapping, which is a resampling technique repeatedly sampling an observed dataset to create a large number of a bootstrap dataset can help better estimate statistics on a population, and was employed. 2000 bootstrap datasets were generated with replacement from the original dataset (n=127). The AUC for both the clinical model (model 2) and clinical model with *TNF* (model 2.2) were generated for each bootstrap sample. Mean and 95% CI of bootstrap-AUC were retrieved with 0.8708 (0.8693, 0.8723) and 0.9149 (0.9137, 0.9162) for model 2 and 2.2, respectively. The absence of overlap between the confidence interval of two distributions of AUCs suggests a significant difference between the 2 AUCs. Paired T-test also confirmed the significant difference ($p < 2.2 \times 10^{-16}$) with an 95% CI of difference in means of -0.0451 to -0.0433. The significant difference between the clinical model with and without *TNF* was also observed in ACPA negative patients (model 4 : mean-AUC 0.9236 (0.9222, 0.9251) and model 4.2 0.9625 (0.9615, 0.9634)). Paired T-test also confirmed the significant difference ($p < 2.2 \times 10^{-16}$, 95% CI -0.0398 to -0.0379).

Overall, introducing the *TNF* methylation to the other clinical variable helped improve the classification performance for both the overall patient group and the ACPA negative group. This suggested *TNF* qMSP assay has added value and the potential to be used as a diagnostic biomarker.

Overall patient

	Clinical model 2 **Model 2**	Clincal model 2 +TNF **Model 2.2**
AUROC	0.8607 (0.7914, 0.93)	0.9043 (0.8489, 0.9597)
R^2	0.497	0.61
At 80% specificity:		
cut off	0.5457	0.4934
Specificity	80.0 (66.3, 89.9)	80.0 (66.3, 89.9)
Sensitivity	80.3 (67.6, 89.8)	85.7 (73.8, 93.6)
PPV	81.8 (68.9, 90.6)	82.7 (70.2, 92.1)
NPV	78.4 (64.9, 89.1)	83.3 (70.1, 91.8)
Accuracy	80.2 (71.3, 87.3)	83.0 (74.5, 89.6)

Confussion matrix	Reference		Reference	
Prediction	RA	Non-RA	RA	Non-RA
RA	45	10	48	10
Non-RA	11	40	8	40



ACPA negative patient

	Clinical model 4 **Model 4**	Clincal model 4 +TNF **Model 4.2**
AUROC	0.9138 (0.8459, 0.9817)	0.9607 (0.919,1)
R^2	0.648	0.757
At 80% specificity:		
cut off	0.2826	0.1629
Specificity	79.6 (65.7, 89.8)	79.6 (65.7,89.8)
Sensitivity	81.5 (61.9, 93.7)	96.3 (81, 99.9)
PPV	68.8 (51.9, 88.1)	72.2 (56.0, 99.1)
NPV	88.6 (74.2, 94.6)	97.4 (86.2, 98.9)
Accuracy	80.3 (69.5, 88.5)	85.5 (75.6, 92.6)

Confussion matrix	Reference		Reference	
Prediction	RA	Non-RA	RA	Non-RA
RA	22	10	26	10
Non-RA	5	39	1	39

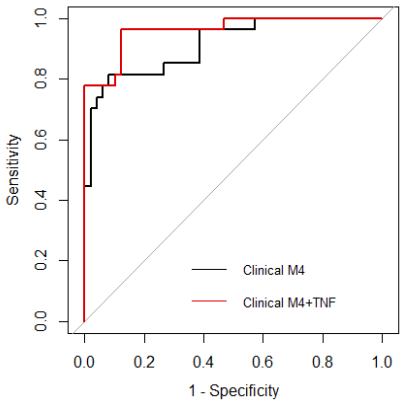


Figure 5-35 The RA classification model and its performance. The clinical model with and without *TNF* overall performance and performance at cut-off giving 80% specificity were illustrated for the overall patient group and ACPA negative patient group. ROC curve of the clinical model without (black line) and with *TNF* qMSP(red line) shows in the right.

5.3.4.3 Validation of RA vs Non-RA classification performance of *TNF* qMSP in a replication cohort.

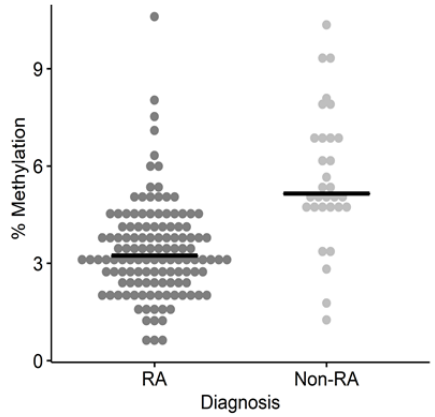
Measuring the RA/Non-RA classification performance of the *TNF* qMSP model using the data used to train model (“in-sample fit”) might lead to over-optimistic results. To further confirmed the RA/Non-RA classification performance of *TNF* qMSP assay, I therefore performed the experiments and analysis in more samples, referred to as a RADAR replication group. This group recruited between 2015-2018, was selected from the RADAR register which is a continuation of the IACON register. However, clinical practice had evolved over time and this register was only included patient with clear IA symptoms while most alternative diagnosis were no longer consented (hence the proportion of non-RA patients was smaller).

I applied the trained *TNF* qMSP model as well as the TNF methylation cut-off level obtained from the discovery cohort analysis and applied it to the replication cohort. This cohort recruited 157 patients with a PBMC samples from the RADAR cohort (continuation of the IACON register) including 126 RA but only 31 non-RA arthritis (mainly PsA and UA) as the inclusion criteria in this more recent register excluded better patients with profiles suggesting non-persistence and other diagnoses (based on clinical experience rather than any particular criteria).

The significantly lower *TNF* methylation in RA patient compared to Non-RA group was confirmed in this dataset (Δ-DM = -1.91%, p=8.36 $x10^{-9}$, Figure 5-36) although slightly lower than in the discovery cohort (Δ-DM= -3.03%, p= $4.1x10^{-9}$). Unadjusted logistic regression analysis to determine the relationship between the diagnosis of being RA with *TNF* qMSP data confirmed their relationship with OR 0.49 (95%CI: 0.36 - 0.64, p= 9.04×10^{-7}). The classification performance as presented by AUROC 0.165 (0.071 - 0.260) was slightly better compared to the discovery cohort.

Using the *TNF* methylation cut-off (4.49%) obtained from the first analysis (discovery cohort) to discriminate high and low risk for RA, confirmed good classification of *TNF* assay as showed in confusion matrix (Figure 5-36,C). The specificity, sensitivity, and accuracy are over than 80% confirming the good prediction of *TNF* qMSP for RA.

A **TNF qMSP assay**



	Medain (95%CI)
RA (n=126)	3.24 (2.52, 4.10)
Non-RA (n=31)	5.15 (4.81, 6.89)

MWU test :

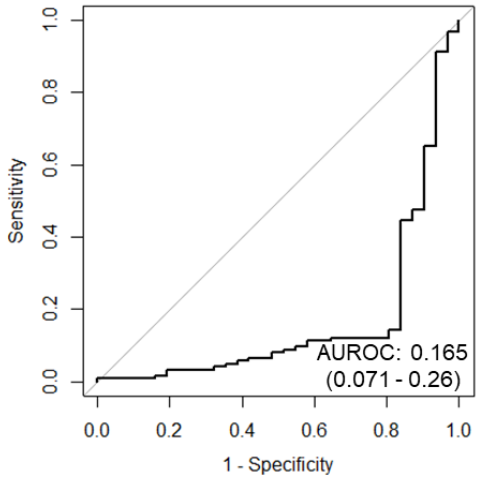
p-value 8.36×10^{-9}

Logistic Regression:

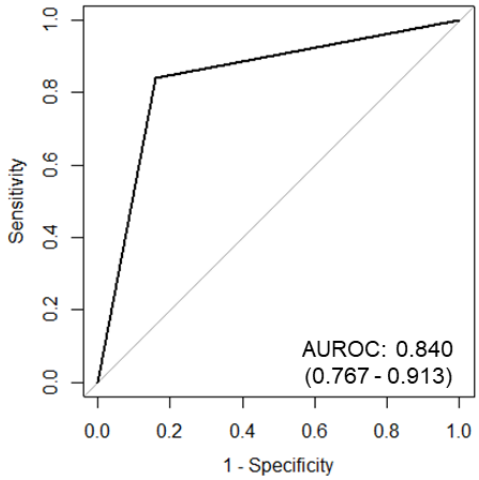
OR 0.49 (0.36, 0.64)

p-value 9.04×10^{-7}

B **TNF qMSP assay ROC**



D **Risk category ROC**



C

TNF qMSP assay (risk category)

Cut-off value

from previous analysis : 4.49 %

High risk/ Low risk 111 / 46

Confusion matrix:

	Actual values	
Predicted values	RA	Non-RA
High-risk	106	5
low-risk	20	26
Specificity	83.9 (66.3, 94.5)	
Sensitivity	84.1 (76.5, 90.0)	
PPV	95.5 (89.8, 98.5)	
NPV	56.5 (41.1, 71.1)	
Accuracy	84.1 (77.4, 89.4)	

Figure 5-36 Univariate predictive value of the *TNF* qMSP assay for overall patients in replication cohort.

A) qMSP *TNF* methylation data for the RA and Non-RA group. Crossbar presents median of methylation level (%). The statistical analysis compares the difference between groups was performed using the MWU

test. The association between *TNF* methylation and the classification of RA was determined using unadjusted logistic regression determining an odd ratio (OR)

B) The overall performance of the *TNF* qMSP levels was determined by an AUCROC (95% CI) analysis.

C) TNF methylation risk category defined as high/row risk of RA using a *TNF* cut-off obtaining from the previous trained model (in discovery cohort).. The confusion matrix of the classification results compared to the actual diagnosis result is displayed. Sensitivity, specificity, NPV, and PPV are described in the table.

D) The overall performance of the *TNF* qMSP risk categorization was determine by an AUCROC (95% CI) analysis.

The added-value of *TNF* qMSP assay on top of other clinical variables on the classification performance for this replication cohort could unfortunately not be performed. I am unable to obtain a complete demographic and clinical dataset form these patients due to COVID situation, and the access to the patient database being restricted, I could not retrieve it personally either. There is currently over 50% missing data for most clinical variable (as shown in Table 5-17) which prevent me to perform further analysis. It is also unlikely to be appropriate to use any method for missing data imputation for this large number. Further analysis after obtaining as much data as could be retrieved from NHS servers is included in the future plan notably to allow me to submit this work to the journal for publication in the few month after submitting my thesis.

Table 5-17 Demographic and clinical parameters of the replication cohort. The descriptive statistic (median (IQR) or frequency) and statistical significant (MWU or X^2 tests) comparison between RA and non-RA group were presented.

Variable	Non-RA n= 31	RA n= 126
Age	52 (35.5, 61.5)	57.5 (49.3, 68.0)
Gender (M/F)	11/20	40/86
Smoking (never/smoker)	14/17	49/75 [2 NA]
RF (positive/negative)	2/14 [15 NA]	50/38 [38 NA]
ACPA (positive/negative)	3/14 [14 NA]	50/36 [40 NA]
Duration	6 (3.7, 8.6) [4 NA]	6.4 (3.7, 11.9) [7 NA]
Tender joint count	5.5 (2.8, 13) [11 NA]	7 (2, 15) [25 NA]
Swollen joint count	3.5 (1, 4.25) [11 NA]	4 (2, 9) [26 NA]
CRP	0 (0, 5) [26 NA]	5 (25, 44.8) [98 NA]
DAS28	4.2 (4, 4.5) [28 NA]	4.4 (3.4, 5.7) [102 NA]

[number NA] indicated the missing data (number of individual).

5.3.4.4 Validation of RA vs Non-RA classification performance of *TNF* qMSP assay using bootstrapping for optimism correction

In any statistical analysis, there is always a risk of model overfitting. The optimism can be estimated by comparing the model performance obtained from the data itself (using the same dataset to develop the model and obtain an AUC) to an internal/external validation (using a difference dataset to test model). This is what I did initially between the discovery and the validation cohort. Although both samples set were small, this produced similar results.

In the absence of a second cohort, an alternative is to create a large number of the datasets using the original cohort by bootstrapping. This can be compared to the original dataset (here combining both discover and replication cohorts). I therefore performed such analysis, which allowed me to calculate a correction for optimism using the approach described in (388-390) as displayed Figure 5-37, A).

The combination of both datasets (discovery and the replication cohorts) was used in this analysis, here referred to as the original qMSP data (n=284, included 190 RA and 94 non-RA). The significantly lower *TNF* methylation in RA patients compared to the non-RA group was confirmed in this dataset (Δ-DM = -2.9%, $p=1.96x10^{-19}$, Figure 5-38,A).

The analysis was then developed stepwise : Figure 5-37,A

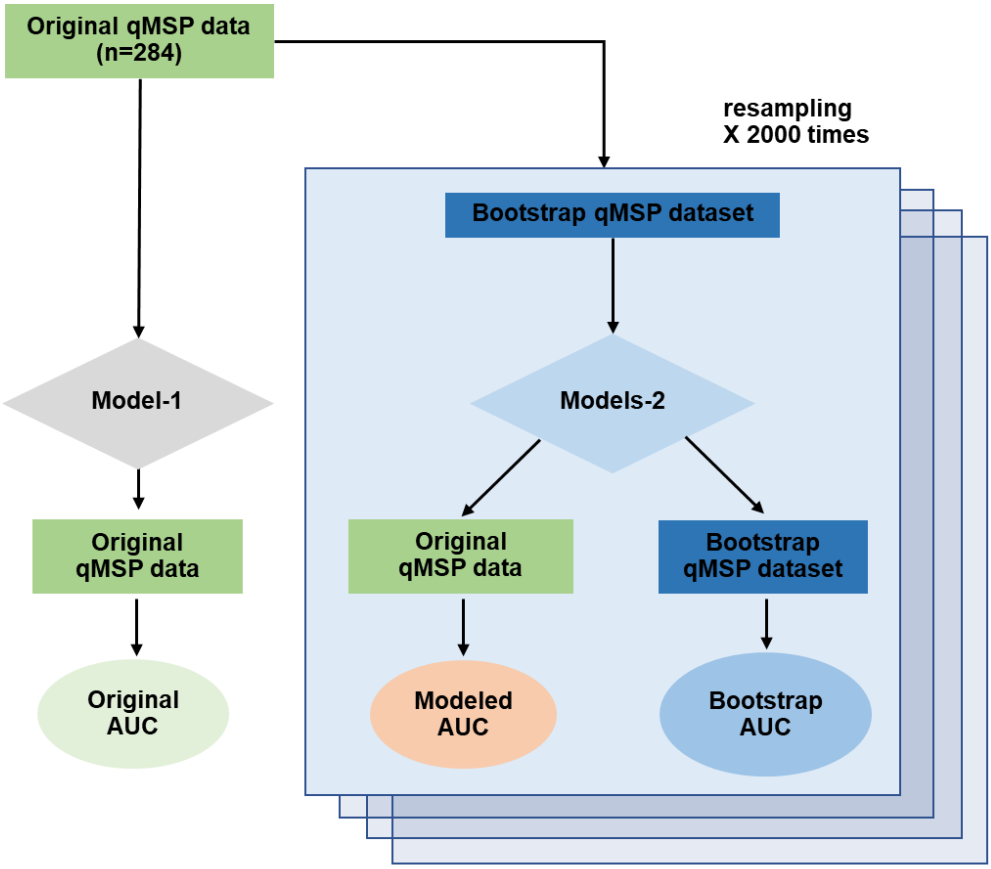
1. *A TNF* qMSP model-1 was developed from the original qMSP dataset and the same dataset was applied to the model to generate an AUC, referred to as the **original-AUC**, which was calculated at 0.171221.
2. 2000 bootstrap dataset were generated.
3. Each bootstrap dataset was used to develop a 2nd model. This model-2 performance was tested using:
    a. The original qMSP dataset to generate a **modeled-AUC**
    b. The same bootstrap dataset to generate a **bootstrap-AUC**.
4. The optimism was calculated as the average of 2000 differences between the modeled-AUC and bootstrap-AUC.

In this analysis, optimism was obtained at -0.000209 (Figure 5-37,B). The optimism corrected AUC was then calculated at AUC of = 0.171221-(-0.000209) = 0.1714301.

After optimism correction, the AUC of *TNF* qMSP model was therefore lower by 0.000209 compared to the apparent-AUC, suggesting a small drop in the classification performance, but still a very good classification model.

Overall, the validation of RA classification of *TNF* qMSP assay either in the validation cohort or using bootstrapping and optimism correction using the combination cohort, confirmed good performance of this assay and showed promising results for the further applicability of this assay.

A Bootstrap approach for optimism correction



B Optimism correction result

Optimism = average of (Bootstrap AUC –Modeled AUC) =-0.000209

Optimism corrected AUROC = Original AUC – optimism =0.1714301

Figure 5-37 Optimism correction using bootstrapping approach:

A) Bootstrap approach for optimism correction.
B) Optimism correction result.

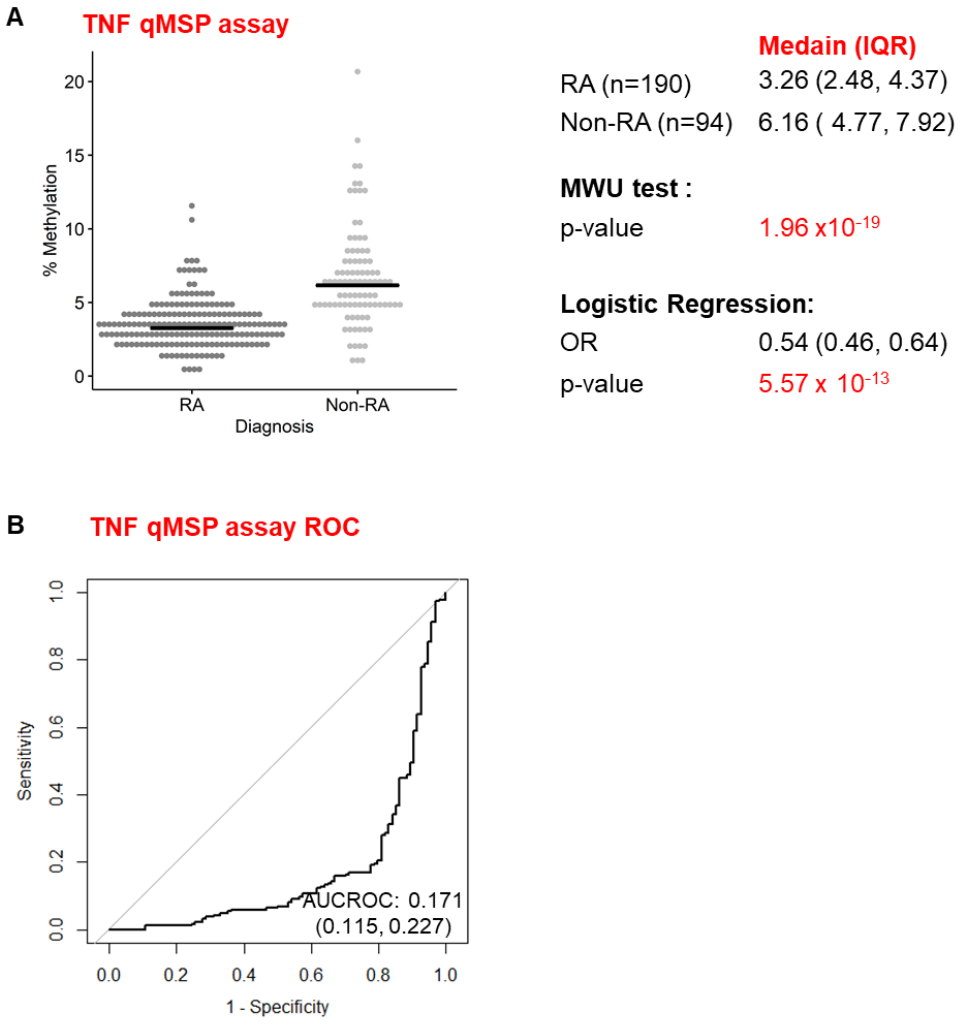


Figure 5-38 Univariate predictive value of the *TNF* qMSP assay in the combination of discovery and replication cohort.

A) qMSP *TNF* methylation data for the RA and Non-RA group. Crossbar presents median of methylation level (%). The statistical analysis compares the difference between groups was performed using the MWU test. The association between *TNF* methylation and the classification of RA was determined using unadjusted logistic regression determining an odd ratio (OR)

B) The overall performance of the *TNF* qMSP levels was determined by an AUCROC (95% CI) analysis (Original-AUC).

5.3.5 Performance of the qMSP assays as a marker for MTX response performed on patients samples

I further performed an analysis to see whether *TNF* and *HDAC4* qMSP could be extended to the prediction of early RA patient's response to MTX.

MTX is the first-line drug in the early treatment of RA. MTX is known to inhibit methionine S-adenosyltransferase (MAT), therefore theoretically reducing the availability of the methyl donor, S-adenosyl methionine (SAM), used for the DNA methylation by DNMT (234, 391-393). Altogether, with the likely effect to induce DNA de-methylation, there is however mix evidence showing both increasing and inhibitory effect of MTX on DNA methylation depending on studies. High-dose MTX in cancer treatment was reported to decrease SAM and increase in global methylation (394). Low-dose of MTX for RA treatment was reported to result in demethylation of *FoxP3* gene in Treg (234) without association with outcome of treatment, but on the other hand, to reverse global DNA hypo-methylation (225, 235).

In RA patients using WB, Nair and colleagues showed 2 CpG sites with change in DNA methylation at 4 weeks after MTX treatment, correlating with response to MTX treatment (395).Another study of leukocyte in early RA treated with MTX observed that changes in global DNA methylation were not associated with MTX response over 3 months but higher baseline global DNA methylation was associated with poor response (396). Therefore, there may be loci-specific versus global DNA methylation effect of MTX that are not yet fully understood.

Some patients who receive MTX respond well and achieve remission (defined by DAS28<2.6 at 6 month post treatment initiation) while some do not respond to this drug (397). This has important consequences for patients as delaying efficacious control of RA results in prolonged inflammation and lead to long-term disability and premature mortality. To predict having a good response for the first-line drug is very important for the patients. Therefore the marker for predicting the response to MTX is of necessity

This analysis aimed to investigate whether *TNF* (or *HDAC4*) gene methylation analysis could be associated with differential response to MTX (comparing baseline and 6 month samples) and whether the qMSP assays for these two genes could be used to predict the patient's response to MTX and be developed as a prognostic or stratification biomarker to tailor the use of MTX only in patients who may clearly benefit from it.

32 RA patients who received MTX (from the discovery cohort) were examined based on having available clinical data for the 6 months following treatment initiation. Of these patients, 15 achieved remission at 6 months and 17 were still presenting active disease (non-remission). PBMC were collected at baseline and week 24, and PBMC DNA was used in the *TNF* and *HDAC4* qMSP assays.

Overall, *TNF* methylation levels at week 24 were significantly lower than at baseline (Δ-DM= -0.91%, Wilcoxon Signed-Ranks test, p=0.0026) (Figure 5-39, A), although this was not observed in all patients. For the *HDAC4* assay, the methylation levels were also significantly reduced at week 24 (Δ-DM= - 2.51%, Wilcoxon Signed-Ranks test, p=0.0498) (Figure 5-39, B).

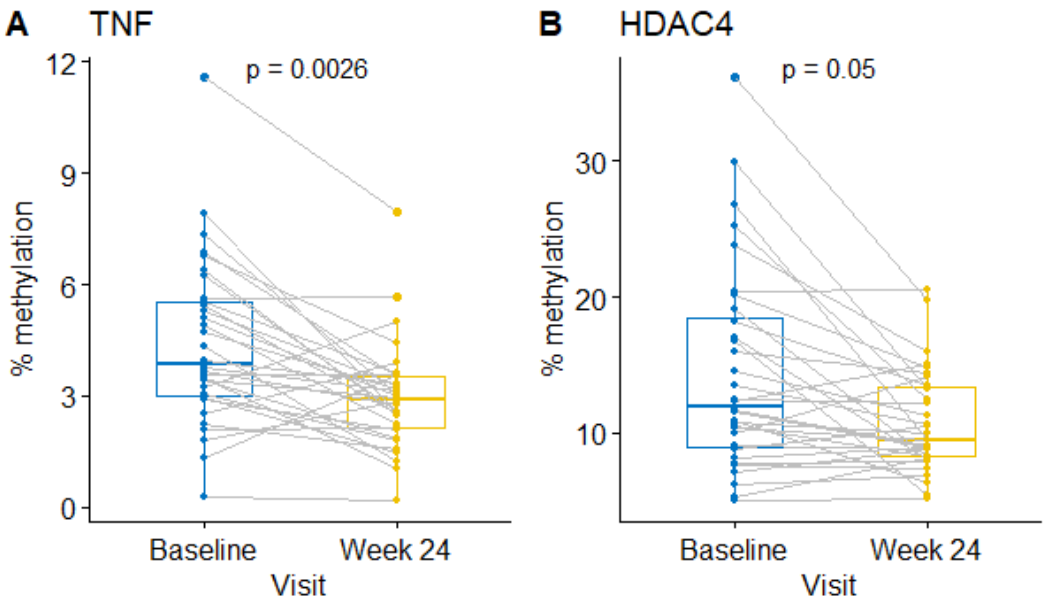


Figure 5-39 Box plot of the levels of DNA methylation (%) for A) *TNF* and B) *HDAC4* genes using the qMSP assays at baseline and week 24. Statistical analyses for paired samples were performed using the Wilcoxon Signed-Ranks test.

Considering each MTX response group individually (Figure 5-40,A, Table **5**-**18**), the methylation dynamic between baseline to week 24, did not show any obvious difference. Both groups showed significant reduction in *TNF* methylation levels at week 24 compared to baseline (∆-DM= -2.25%, p=0.0103 for remission and ∆-DM= -1.19%, p=0.0150 for non-remission).

The dynamic of the methylation of *HDAC4* in the two MTX response group also showed no difference (Figure 5-40, B). *HDAC4* methylation of both groups was significantly lower at week 24. (∆-DM= -2.2%, p=0.0302 for remission and ∆-DM= -4.3%, p=0.0150 for non-remission).

Therefore, despite disease remission, the loss of methylation in the *TNF* and *HDAC4* genes continues degrading over 6 months, suggesting that *TNF* DNA methylation might not be directly related to the mechanism of response for this drug.

However, at baseline these data also suggested higher % of methylation of the *TNF* gene in patients achieving remission (Figure 5-41 and Table **5-19**, n=20, 4.71 ± 2.38 % of methylation) compared to those who do not (n=21, 3.26 ± 1.33 %, p=0.0184), while this was not the case for *HDAC4*. This suggested the potential for an association between baseline levels of *TNF* methylation and response to MTX.

A binary logistic regression was performed. Higher baseline *TNF* methylation (%) show association with remission (OR = 1.59, 95% CI: 1.09 - 2.55, p =0.0319) (Figure 5-42). The AUROC was 0.714 (95% CI: 0.553 - 0.876) suggesting the good classifying performance of the *TNF* qMSP. At a chosen cut-off for an 80% specificity set at 4.65% of methylation, descriptive performance of the risk categories for remission/non-remission were calculated for specificity, PPV, NPV (displayed in Figure 5-42). PPV and NPV were 62.9% and 71.4%, respectively. Altogether this suggest good discrimination and 65.9% of individual in the groups being correctly predicted.

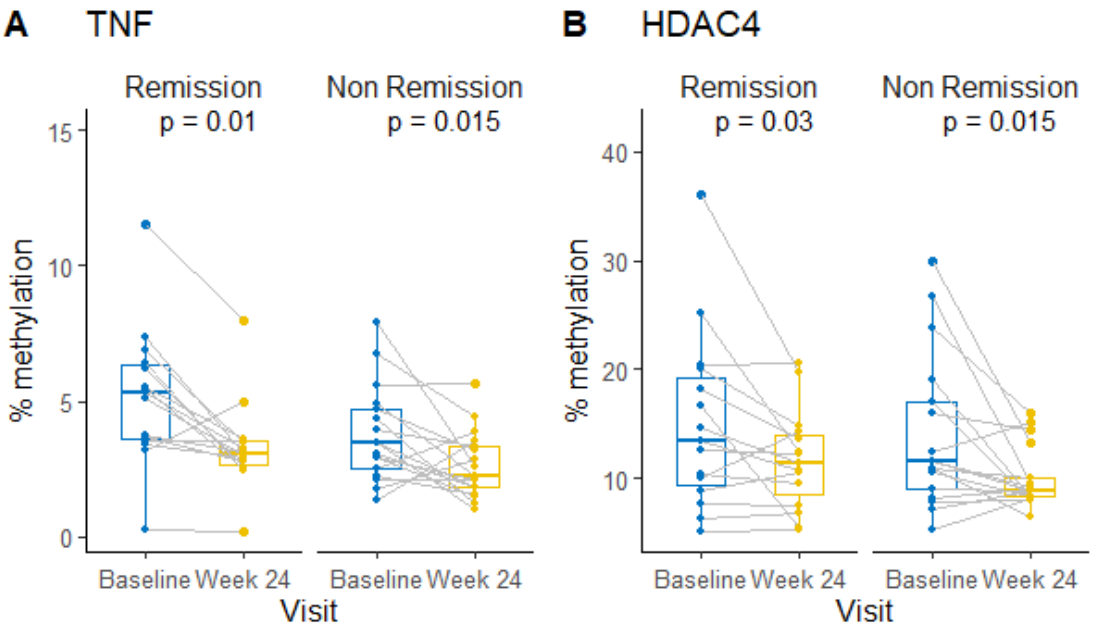


Figure 5-40 Box plot of the % of methylation in remission and non-remission groups at different visiting point for the A)*TNF* and B) *HDAC4* qMSP assays. Statistical analyses for paired samples were performed using the Wilcoxon Signed-Ranks test comparing the methylation between baseline and week 24 of each MTX response group.

Table 5-18 **Descriptive statistic of *TNF* methylation of remission and non-remission group at baseline and week 24.**

Visit	MTX response	n	Methylation (%)		
			Mean	SD	Median (IQR)
Baseline	Remission	15	5.18	2.52	5.30 (3.62, 6.31)
Week 24	Remission	15	3.31	1.63	3.05 (2.68, 3.54)
Baseline	Non-remission	17	3.77	1.78	3.45 (2.54, 4.69)
Week 24	Non-remission	17	2.64	1.24	2.26 (1.82,3.32)

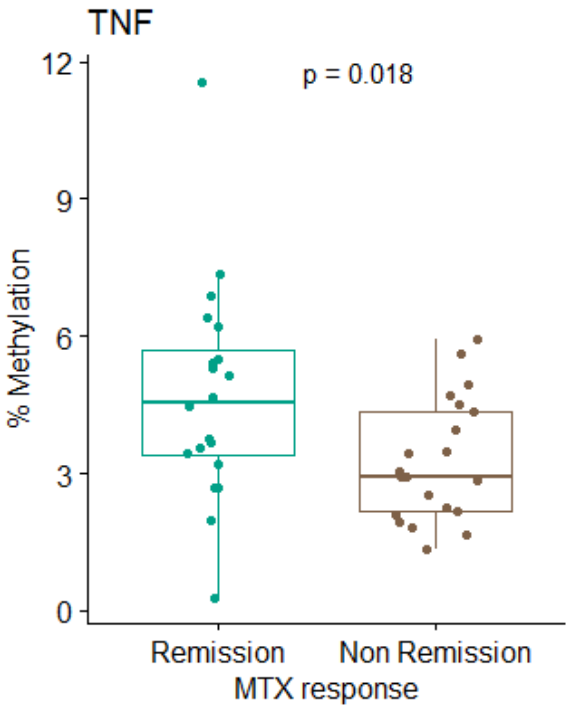


Figure 5-41 Boxplot of DNA methylation levels (%) of the *TNF* genes at baseline, in remission and non-remission group. Statistical analyses were performed using the Mann-Whitney U test comparing the methylation between remission and non-remission group.

Table 5-19 Descriptive statistic of *TNF* methylation of remission and non-remission group at baseline.

Visit	MTX response	n	Methylation (%)		
			Mean	SD	Median (IQR)
Baseline	Remission	20	4.71	2.38	4.65 (3.38, 5.69)
Baseline	Non-remission	21	3.26	1.33	2.93 (2.16, 4.35)

	TNF (qMSP)
OR	1.59 (1.09, 2.55)
p-value	0.0319
AUROC	0.714 (0.553, 0.876)
At 80% specificity:	
Probability cut off	0.575
Value at cut off	4.65%
Specificity	80.9 (58.1 - 94.6)
Sensitivity	50.0 (27.2 - 72.8)
PPV	62.9 (42.4 - 80.6)
NPV	71.4 (41.9 - 91.6)
Accuracy	65.9 (49.4 - 79.6)

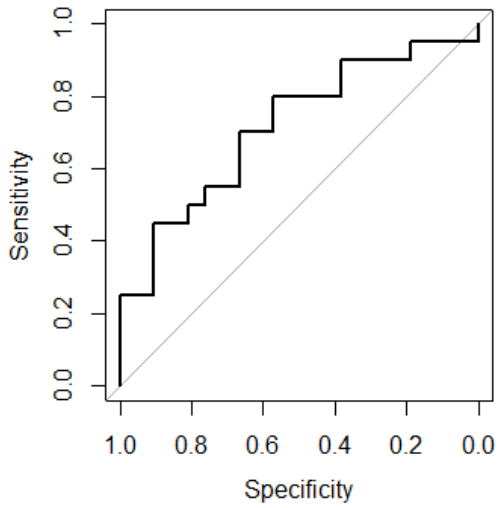


Figure 5-42 *TNF* qMSP logistic regression. The table presents the overall performance of the models and performance at cut-off giving 80% specificity. ROC curve of *TNF* qMSP shows in the right.

Although *TNF* qMSP alone shows promising potential to predicting the MTX response, it needs to be evaluated for added value compared to other demographic and clinical parameters.

I first explored the relationship between the response to MTX and the individual demographic and clinical parameters including age, gender, RF, ACPA, TJC, SJC, CRP, DAS, and symptoms duration to the using binary logistic regression. Unfortunately, none of these parameters shows any trend for significant association with MTX response (OR 95% CI within 1, $p>0.05$) as this is far too small a group to be able to detect any association (descriptive statistic of all parameters between remission and non-remission group and the value from logistic regression can be found in Table 5-20). However, it also aligns with current literature, showing no particularly strong association of any demographic or clinical parameter at baseline with response of MTX. The fact that the *TNF* qMSP data showed association suggest they may be quite a powerful tool if able to show prediction in this group.

Again, a replication cohort was intended (using the RA patients of the replication cohort treated with MTX) but no data could be retrieved in time to allow this analysis to be performed, while it is planned in the future to allow me to publish this data.

Overall, this analysis showed no difference in the continued reduction of the *TNF* gene methylation from baseline to week 24 between the remission and non-remission patient treated with MTX. At baseline, differences in *TNF* methylation levels between the 2 groups (despite small number n=41) were clearly observed. An unadjusted Logistic regression analysis showed predictive value for the TNF methylation levels at baseline and MTX response, with quite high and promising performances that needs replication in more samples.

A replication of this work in the RADRA cohort has not been possible due to lack of follow-up during the COVID pandemic and the lack of data to access MTX treatment outcome at 6 month.

Table 5-20 Descriptive statistic and logistic regression of MTX response and Demographic and clinical parameters. The descriptive statistic (median (IQR) or frequency) and statistical significant (Mann-Whitney U test or chi-square test) comparison between remission and non-remission group were presented.

Independent Variable	Descriptive statistic			Regression	
	Remission n= 20	Non-Remission n= 21	p-value	OR (95% CI)	p-value
Age	61 (42.5, 72)	52 (46, 64)	0.6385	1.01 (0.97 - 1.05)	0.5633
Gender (M/F)	6/14	3/18	0.4022	1.90 (0.50 - 8.62)	0.3806
Smoking (never/smoker)	6/14	7/14	1.0000	1.17 (0.31 - 4.48)	0.8187
RF (positive/negative)	6/14	10/10*	0.3329	0.42 (0.11 - 1.53)	0.2010
ACPA (positive/negative)	8/12	9/12	1.0000	0.89 (0.25 - 3.10)	0.8528
Symtom duration	5 (3.8 - 9)	5 (2 - 12)	0.7631	0.922 (0.81 - 1.04)	0.1966
Tender joint count	7.5 (2.8 - 12.3)	11 (9 - 14)	0.1778	0.97 (0.88 - 1.05)	0.4258
Swollen joint count	1.5 (0.8 - 7.5)	6 (3 - 9)	0.1034	0.93 (0.82 - 1.04)	0.2775
CRP	7.5 (0 - 19.3)	12 (5 - 21)	0.4594	0.99 (0.96 - 1.01)	0.3851
DAS	4.2 (3.3 - 5.2)	5.4 (4.2 - 5.9)	0.0296	0.62 (0.35 - 0.99)	0.0605

5.4 Biomarker development summary/discussion

An epigenetic DNA methylation biomarker assay for RA classification was developed through the biomarker development workflow from target identification, target verification, assay development, and assay validation steps.

3 strategies for target identification were designed with different benefits and limitations but all shared a common focus on RA specificity, based on the hypothesis that such change in methylation occurred early in RA. Of all candidate CpGs, 8 top candidates (associated with the *TNF, IFITM1, RPTOR, ATP6V1H, IRF8, HDAC4, MIR21,* and *PSMB9* genes) which show more potential as diagnostic biomarker were highlighted.

A target verification step to verify the methylation status and ensure specificity and suitability for assay design, was performed using bisulfite sequencing. The verification for *IFITM1* unfortunately failed. This pointed to the weakness of the target selection strategy-1 and the need for more DNA methylation data resource form other cells types. (which was part of the design of strategy 2 and 3). However, due to the tedious process of bisulfite conversion optimisation of each target sequence and the limitation of my study time, the verification for other genes (except from *TNF* which had already been sequenced to confirm its methylation status) was skipped and continue directly using the qMSP assay development process.

Several qMSP assays were developed to detect the DNA methylation status of the target sequence which include the candidate CpGs and it's surroundings (where primers bind). Assay development was attempted for 5 genes (*TNF, HDAC4, PRTOR, MIR21, and IRF8*) by either SYBR green or TaqMan based qMSP while primer design for the other two (*ARP6V1H and PSMB9*) was unsuccessful due to the sequence around the CpG location being inadequate for a qPCR reaction. the *TNF* and *HDAC4* qMSP assays (using TaqMan-based detection) were successfully developed with good assay efficiency and sensitivity.

For the biomarker assay value, the *TNF* and *HDAC4* qMSP assays were tested with clinical samples for their performance in RA classification; to differentiate between RA and other early arthritis.

The *TNF* and *HDAC4* methylation levels quantified by the qMSP assay showed a lower percentage of methylation readings compared to the methylation detected by the Illumina array and bisulfite sequencing. This might be because the calibration of my qMSP assay was based on the relative fold of methylation

of the sample normalised to the levels observed in the 100% methylated DNA control. Theoretically, when an equal amount of control DNA (100% methylated DNA control) is compared to the test sample, the methylation levels in the sample is expressed as a relative methylation change compared to the standard (100% methylation control). However, in my experiment, the 100% methylation DNA control was commercially obtained (although several batches were used at a defined concentrations) as an already bisulfite converted template, while the test sample was bisulfide converted on-site. After bisulfite conversion, the DNA quality and quantity may be less due to the nature of the bisulfte conversion process (note that measuring bisulfite converted DNA concentration using nanodrop may also not give as a fully reliable result now that the DNA sequence is composed of 5 bases rather than 4). As the initial amount of template for the reference reaction (100% DNA control) and the sample reactions are unlikely to be absolutely equal, the value obtained following the calibration (as detailed in method part) may have been affected slightly by the overall process.

Nevertheless, the *TNF* qMSP assay showed good classification performance in PBMC DNA sample in a discovery cohort (for both the overall patient cohort n= 127 and for the ACPA negative patients n= 92). It also showed added-value in classification for the *TNF* qMSP assay when compared with current clinical models. On the other hand, the *HDAC4* qMSP assay showed differences in methylation between RA and non-RA in PBMC DNA but was not specific enough to improve RA classification. This result was unexpected as *HDAC4* was a candidate selected by strategy 3 which emphasised the detection of Δβ of RA in PBMC. Besides, 3 proximal CpGs associated with *HDAC4* were on the candidate list. The failure of *HDAC4* qMSP assay in clinical samples points to;

(i) some weakness in the target identification steps, which did not include sufficient data resources in other **early** form of arthritis in the different cell types. This issue was difficult to address due to the limited availability of data resource but points to the need for datasets **relevant** to the question in order to design successful assays (i.e. early stage of disease, other early inflammatory disease related to RA classification, all cell types) which could be leveraged with more reachable/affordable array technology, and more tool for the analysing of DNA methylation across platforms, better target selecting strategies that are fitted to the research question.

(ii) the importance of the target verification step. DNA methylation data of candidate CpG retrieve form the array (and the immediate surrounding

CpGs which do not appear in the array data as spaced) need to be confirmed as they all affect primer/probe design and binding. Similar methylation status of the CpGs surrounding the candidate is important (discrepancy/variability in methylation status of the overall region will indeed prevent assay design). A careful verification of the candidate CpG/surrounding region before developing the qMSP assay would help this issue.

TNF qMSP classification performance was also confirmed in the replication cohort (n= 157). Validation of this assay's performances using bootstrapping and optimism correction also confirmed its good AUC. This established that the *TNF*qMSP model developed in a discovery cohort had a good classification performance and can be used for the prediction of RA diagnosis in a large sample. It has to be noted that the samples of the discovery and replication cohort selected for this project have a similar background since both cohorts recruited samples from patients who attended the EAC at Chapel Allerton Hospital in Leeds, UK, however at 2 different points in time and with an evolution of clinical practice between the 2 groups. To translate this assay into the clinical practice, hence applying it to various population background (taking into account ethnicity for example as well as recruitment capacities differing between health care system of different countries), more data would be needed for developing/training the prediction model, determination of the suitable cut-off value, as well as more validation. In addition, as DNA methylation can be influenced by the environment, distinct populations may have different methylation baselines. Therefore re-defining local/country specific cut-offs for this biomarker might be necessary for each population.

The use of *TNF* qMSP assay was also extend from classification to the prediction of MTX response. The small cohort I tested (n= 41), showed that *TNF* methylation of PBMC DNA detected at baseline could predict the response of MTX treatment. This was a promising result, however, it needs more validation in the larger cohort which includes the future plan. This was also an unexpected result (although it was also observed for the IL17A assay by my supervisor (378), as the CpG chosen was not selected with this clinical question in mind, which may have needed another type of strategy. This is further discussed in my general discussion.

Chapter 6 General Discussion

My thesis was developed around 2 main aims. My findings can be summarised as follows:

- Part1 aimed to gain more understanding of the early molecular and cellular events in the RA disease pathogenesis. The analysis of Illumina 450K genome-wide array form naïve or memory T-cell CD4+T-cells and monocyte of early RA patients demonstrated that methylation change occurred at an early stage in RA and display specific patterns for each cell type. Naïve CD4+T-cells are the most susceptible to such modifications although DNA methylation alterations points to a role for IL6 signalling in early RA pathogenesis in the 3 cell types. This has a particular central role in naïve CD4+T-cells, with diversification towards other pathways (notably TNF, IFN-signalling, Th17 differentiation) very early in the disease course. My findings allow to propose DNA methylation as a mechanism for the model of IL6 induced T-cell differentiation previously observed in RA by my supervisor.

- Part2 aimed to select potential candidates for further development as a biomarker for diagnostic in early RA. The selection of candidate CpG was based on having a methylation pattern that is specific to CD4+T-cells, to RA versus other IA that is detectable in an easily accessible sample (i.e. blood). The technology chosen was a qMSP assay. An assay measuring change in the TNF gene methylation by qMSP was successfully developed to work with PBMC DNA sample. The assay could differentiate between RA and other non-RA IA patients referred to an EAC. A regression analysis showed that this assay itself has a good classification performance but importantly also has added-value as a diagnostic biomarker in a classification model including other demographic, risk factors and clinical variables. This *TNF* qMSP assay also shows a potential to be used as a marker to predict the response to MTX using the DNA sample at baseline, differentiating responders from non-response with a good classification performance.

The study of epigenetic modifications which has been proposed as an early event associated with the development of certain diseases (187, 210, 215) before other molecular mechanism (e.g. gene expression, protein production) and clinical manifestations is very useful for both gaining a deeper understanding of disease pathogenesis and developing biomarkers. (238, 240, 312, 380). DNA methylation is an important modification that shapes the chromatin and is the focus of my project. In a study of methylation alteration associated with early (drug naïve) disease pathogenesis, the question that is often addressed is whether methylation changes are a cause or a consequence of the disease process. There is still no valid answer to this question caution was raised about interpreting result of epigenetic associations with disease (238, 398). RA is a disease which like many others has been associated with many epigenetic modifications (190). Since the beginning of my PhD several further pieces of evidence have supported that epigenetic modifications is important causative in early RA and not a consequence of long lasting disease duration (or medication) (238, 240).

The Methylation change I have identified occurred at the early stage of the disease in naïve/memory CD4+T-cells and monocytes in RA patients. Several other studies using SF, T-cell, and B-cell (238-240) have also been reported in early RA. In my view, study methylation changes at the early event in the drug naïve sample could lead to the pathway that contributes to disease development or at least could help understand the event at a closer look to the origin. It is beneficial to both pathology and biomarker study.

Methylation changes were detected in peripheral blood cells. RA is an autoimmune disorder that primarily affects joints. In the past, most of the epigenetic studies were focused on cells at the disease site e.g., the joint, using SFs, macrophages, or T-cells. As immune cells circulate around the body and RA is considered a systemic disease, peripheral blood (PB) which is an easily accessible sample, has great potential as a good tissue source with sufficient representability for the overall events in disease pathogenesis. As such many biomarkers were developed for RA that use blood (plasma or cells). A study of methylation changes which define a DNA methylation signature in RA in SF was confirmed in circulating naïve CD4+T-cells suggesting that the blood is suitable as a tissue source of epigenetic signatures. Easily accessible samples can facilitate research (especially in the biomarker field) and the fact that certain events could be detected in both the local and systemic compartments validate the approach I used.

My study observed methylation changes in 3 cell subsets: naïve CD4+T-cells, memory CD4+T-cells and monocytes. The unique pattern of methylation changes in each cell type emphasizes independent possible roles for each cell in RA pathogenesis. Studying cell-specific methylation profile therefore provides more information and is more useful for understanding disease pathology than profiles obtained in WB or PBMC due to the complexity introduced by mixed cells population in the sample. On the other hand, DNA methylation profile of WB or PBMC are more likely to be useful for a biomarker development as it can be translated into a more practical technical test.

Fewer DM genes were observed in memory CD4+T-cells and monocytes compare to naïve CD4+T-cells. The majority of DM genes in memory T-cell showed hypermethylation suggesting a silencing effect by analogy with cancer. Monocytes are cells that respond to stimuli fast with whole program of gene expression but which have a short life span. If methylation alterations are meant to foster the development of the RA disease, it is reasonable to hypothesise that such modification should preferably occur in cell type with a long-life span. Considering that naive CD4+ T-cells are prone to be activated and to differentiate from variable stimuli to various T helper cell or regulatory cell subsets, it is also conceivable that methylation changes occur preferably in this cell type as they may be more sensitive to change than memory cells.

DNA methylation of circulating naïve CD4+T-cell (and also other specific cell types) was previously studied in established RA(est-RA) by another group (312). They analysed DNA methylation from 63 fully established RA and 31 HC in naïve T-cell, memory T-cells, B-cells, and monocytes using the same Illumina 450K array. The difference of methylation profile in est-RA compared to HC only identified 16 DM-CpG in naïve and 1 DM-CpG in memory CD4T+cells, with limited DM in other cell types, Overall, these findings were similar to what I observed in early RA, emphasizing the relevance of naïve CD4+T-cells in RA. The top 10 DM genes in est-RA in naïve CD4+T-cells included *TYK2, PRKAR1B, ABCC4, COMT, CAI2, MCF2L, GALNT9, C7orf50*, however, these showed no overlapping with my DM gene list in early RA patients. The difference in the statistical approach and the stringency of the analysis might be a factor explaining the difference in the result from two studies. However, the methylation dynamics over the time course of disease since its development is also likely to be an important factor. The methylation changes associated with different disease stage are indeed different with evidence of different patterns seen in a study comparing DNA methylation of SF in HC, very early RA (symptom duration less

than 3month) and est-RA (238). That study confirmed significant DM genes in early RA (compare to HC) and also identified DM gene at each disease stage, clearly shown using a PCA analysis. Some of these methylation alterations occurred in important functional pathways that were shared between early and established-RA (cadherin, integrin (cell adhesion) and WNT signalling pathways, components, the actin cytoskeleton and antigen presentation) but some were more dominant at a particular disease stage for example cell apoptosis/anti-apoptosis pathway are more dominant in early RA, while cell adhesion, potassium/calcium transport pathway are more dominant in established-RA.

In my opinion, methylation changes at established stages of disease are less likely to be important for disease progression, while reflecting better consequences of the disease and effects of treatment (and leading to heterogeneity), while methylation change at the early stages (drug naïve) suggest a specific immune cell activation is likely to be important to the establishment of disease and resulting in the activation of relevant signalling cascades.

My study is the first to present DNA methylation of naïve CD4+T-cell in early RA patients. The strength of this study is that the methylation was studied in early RA patients and use specific cell types. My work provided a novel understanding of pathways involved at the beginning of pathogenesis highlighting IL6, TNF, Th17, IRF and allowed me to update the model of dysregulated naïve CD4+T-cells differentiation in RA with epigenetic modifications possibly driven by IL6 as a molecular mechanism (the detail discussed in discussion part 1).

DNA methylation in naïve CD4+T-cells was also studied in other autoimmune conditions such as SLE and primary Sjögren's syndrome (399, 400) using methylation wide array. Methylation changes in SLE (compared to HC) were mainly related to interferon regulated genes e.g., *IFIT1, IFIT3, MX1, STAT1, IFI44L, USP18, TRIM22* and *BST2*. In Primary Sjögren's syndrome, DM also showed to involve interferon regulated genes e.g. *STAT1*, *IFI44L*, *USP18* and *IFITM,* solute carrier proteins, and also *LTA*, which is a protein of the TNF family. Methylation changes in interferon related genes that are dominant in both these diseases was also observed in RA (however amongst other pathways) suggested a pathway common between all 3 autoimmune diseases. Methylation changes in pathways that are unique to RA therefore suggests a DNA methylation signature specific to this particular disease.

Gene not coding for proteins were also included amongst the DM-genes. Micro-RNA (MIR) are important regulator of gene expression. *MIR21* is an interesting gene at the top of MIR showing DM. DM changes were detected in 15 CpG (hypomethylation) associated with *MIR21* (in the array) in naïve cells (Δβ= - 0.26, p-value = 4.49 $x10^{-4}$). *MIR21* could also have a strong association with the disease pathogenesis. Most of the genes targeted by *MIR21* are related to IL6 signalling. Another study showed dysregulation of *MIR21* expression in PBMC and CD4+T-cells in association with an imbalance in Th17 and Treg cells in RA patients. Increased expression of STAT3, one of the key transcription factor for Th17 differentiation was also showed suggesting that the possible involvement of *MIR21* in RA pathogenesis uses STAT3 axis to disturb the balance of Th17/Treg cells (401). Furthermore screening of 750 MIR in pre-clinical RA serum samples (ACPA+ individual with arthralgia) by another group in our department, showed changes in MIR21 expression (both using a MIR array and RT-qPCR) associated with progression to clinical symptoms and the development of RA (402). A summer placement student in the group further confirmed the increased expression of MIR21 in serum samples from early RA patients using qPCR with an added value as a diagnostic biomarker for RA classification.

In the biomarker development part of my Thesis which aimed to develop a RA diagnostic biomarker, the candidate CpG were selected based on showing methylation changes specific for RA. One of the strategies for selecting a candidate planned to use available DNA methylation profiles from other IA to exclude the methylation changes that could be common with RA. Of all the candidates recruited from the various selection strategies, *TNF* was the only gene that passed through all biomarker development steps. The *TNF* qMSP assay developed could be used to accurately classify the patient with progression to RA form other early IA and provide a good value as a diagnostic biomarker.

DNA methylation change in the *TNF* gene was studied in other fields notably cancer, neurodegenerative diseases, diabetes, and obesity. In Colorectal cancer tumour, DNA methylation induced silencing of the TNF gene, associated with the lower expression of *TNF* and showed the possibility that it could act as biomarkers for prognosis and future immunotherapeutic strategies (403). Hypermethylation of the *TNF* gene promoter in PBMC *of* type1 diabetes patients was positively associated with homocysteine metabolism and showed the possibility to be develop as a marker of early risk (404). PBMC *TNF-α* promoter methylation was also reported to be a good inflammatory marker

predicting hypocaloric diet-induced weight loss in overweight patients (405). Despite the potential to be further developed as a biomarker most of *TNF* DNA methylation as a biomarker remained in the discovery phase that still needs future confirmation and validation. No assay has yet been reported and my *TNF* qMSP therefore has the potential to provide a test for much more than only RA.

There is yet no official approval of any qMSP assay in clinical use for RA; and many studies are still in the discovery phase. In my project, I developed one assay following the principle of qMSP. The use of other candidate genes was limited in RA, however, the *IL17* gene is another good example of the potential of qMSP. The *IL17* gene family was shown to be DM in RA patients, as showed by my observations and that was used by my supervisor to initiate a collaboration with a company. This work confirmed the value of a qMSP assay for the *IL17A* gene as diagnostic biomarker in early drug naïve RA and for MTX induced remission (378). DNA methylation level of the *IL17* gene in WB DNA quantified by qMSP was reduced in early RA patient compared to other IA patients. The DM of the *TNF* gene was specific to naive CD4+T-cells while not detected in other cells although some may have contributed to the assay results. As such, my qMSP assay was sufficiently sensitive to detect DM in CD4+T-cell DNA and PBMC but not in WB. This is reflecting the difference between the highly specific epigenetic commitment of the *IL17A* gene in Th17 cells (YES versus NO), compared to variable TNF gene methylation (less versus more). The ability of my assay to differentiate between RA and other arthritis groups nonetheless confirmed the potential of DNA methylation assays to be used as a diagnostic biomarker / predictor of MTX induced remission in RA.

In RA disease management, gaining access to the right treatment is as important as early diagnosis. Identifying biomarkers that predict treatment response in RA is a clinical priority especially for the first-line drug used in RA treatment, MTX. In my project, although the initial design of *TNF* qMSP assay was meant to be developed as a diagnostic biomarker, the *TNF* qMSP was also tested for its ability to predict the response to MTX. This was achieved although my data are only a 1st exploration of this potential. On the other hand, MTX was known to affect DNA methylation itself. In recent research (2019), analysing RA patients EWAS data for good and poor response to MTX at week-0 and week-4 using WB proposed 2 CpG (associate with *RPH3AL* and *WDR27*) with change in DNA methylation between 2 time points for response of MTX (395). As the methylation of these 2 CpG was not different at baseline, it was suggested that the changes in methylation at week-4 were due of the effect of MTX acting differently on patients.

In my work, the assay showed that the *TNF* methylation change over time (from baseline to 6 months after treatment) was not different between the good and poor MTX response group. This is suggesting that methylation of CpGs in *TNF* gene at least, are not be directly affected by MTX but continue to be affected by the disease. However, these two response groups showed difference in the initial levels of *TNF* gene methylation at baseline. Higher *TNF* methylation (which is closer to health) was observed in the group achieving remission. This better status at the starting point might simply indicate patients with a less advanced disease and therefore being more responsive to the treatment. This was not reflected by other clinical markers (joint count, CRP) or demographic (disease duration) and therefore unique to the *TNF* qMSP assay. Higher *TNF* level at baseline therefore showed good value for the prediction of response for MTX treatment in a regression analysis. Further work on additional patient (ongoing) is needed to confirm the predictive value of *TNF* qMSP for MTX response.

The association of response to MTX with baseline methylation level rather than the methylation change over time observed here was also reported in a study of leucocyte DNA methylation in early RA patient (396). They observed that changes in global DNA methylation was not associated with MTX response over 3 months while the higher baseline global DNA methylation was associated with MTX poor response and a smaller reduction in DAS28 (396). The difference in direction of association (high methylation to a good or poor response) between this study and my study may be loci-specific as opposed to a global effect of MTX on DNA methylation that are not yet fully understood.

It would be interesting to see whether my *TNF* qMSP assay could also be applied to monitoring methylation change in other diseases treated with MTX. The study of the response of Osteoarthritis (OA) to MTX is ongoing. OA is considered a low-inflammatory disease with very slow progression, that needs a sensitive biomarker for detecting and monitoring changes that may be introduced by treatment.

Chapter 7 Conclusion and future perspective

Altogether my data confirmed the proposed hypothesis suggesting that methylation change occurred early in RA pathogenesis, preferentially in naïve CD4+T-cells. These changes affect several pathways (mainly IL6/STAT3 linked to TNF-α, IFN signalling genes, and Th17 differentiation) confirming a role for these important physiological pathways toward disease development. The methylation changes also occurred in genes of the epigenetic machinery itself, which in turn could affect further the overall DNA methylation process. My data together with work performed in collaboration with others researchers in the group brought more strength to the overall idea that IL6 induce a form of T-cell differentiation leading to the development of an atypical subset of naïve CD4+T-cells resembling *in vivo* abnormally differentiate cells observed in the past, and contributing to RA pathogenesis.

Understanding more about the mechanism involved in disease pathogenesis helps point to new target/pathways for treatment, especially the IFN-signalling and the Th17 polarisation where potential drugs may be tested at the right time in the right patients. Furthermore, this also suggest that the epigenetic machinery itself may be a good target for prevention strategies in pre-clinical RA.

A *TNF* qMSP assay was successfully developed which showed excellent performance at RA classification. An additional biomarker will be helpful for RA diagnosis, especially in an ACPA negative patients, who might benefit the most from an earlier classification and access to treatment. The *TNF* qMSP assay also shows a promising result for predicting MTX response, which if confirmed, will be beneficial to stratify the right patient for access to right the treatment at the right time towards fulfilling the promises of personalised medicine. This biomarker work provided very strong foundation for further validation in a larger cohort and showed the potential for going into clinical use and improving the management of patients with RA.

List of References

1. Symmons D, Turner G, Webb R, Asten P, Barrett E, Lunt M, et al. The prevalence of rheumatoid arthritis in the United Kingdom: new estimates for a new century. Rheumatology (Oxford). 2002;41(7):793-800.
2. Gibofsky A. Overview of epidemiology, pathophysiology, and diagnosis of rheumatoid arthritis. Am J Manag Care. 2012;18(13 Suppl):S295-302.
3. McInnes IB, Schett G. The pathogenesis of rheumatoid arthritis. The New England journal of medicine. 2011;365(23):2205-19.
4. National-Audit-Office. Services for People With Rheumatoid Arthritis (House of Commons, Report by the Comptroller and Auditor General, Session 2008-2009). The Stationery Office; 2009.
5. Silman AJ, Pearson JE. Epidemiology and genetics of rheumatoid arthritis. Arthritis Res. 2002;4 Suppl 3:S265-72.
6. Gregersen PK, Silver J, Winchester RJ. The shared epitope hypothesis. An approach to understanding the molecular genetics of susceptibility to rheumatoid arthritis. Arthritis Rheum. 1987;30(11):1205-13.
7. Goronzy JJ, Zettl A, Weyand CM. T cell receptor repertoire in rheumatoid arthritis. Int Rev Immunol. 1998;17(5-6):339-63.
8. Allen S, Turner SJ, Bourges D, Gleeson PA, van Driel IR. Shaping the T-cell repertoire in the periphery. Immunol Cell Biol. 2011;89(1):60-9.
9. Roudier J. Association of MHC and rheumatoid arthritis. Association of RA with HLA-DR4: the role of repertoire selection. Arthritis Res. 2000;2(3):217-20.
10. Wucherpfennig KW, Strominger JL. Selective binding of self peptides to disease-associated major histocompatibility complex (MHC) molecules: a mechanism for MHC-linked susceptibility to human autoimmune diseases. J Exp Med. 1995;181(5):1597-601.
11. Hill JA, Wang D, Jevnikar AM, Cairns E, Bell DA. The relationship between predicted peptide-MHC class II affinity and T-cell activation in a HLA-DRbeta1*0401 transgenic mouse model. Arthritis Res Ther. 2003;5(1):R40-8.
12. Schonland SO, Lopez C, Widmann T, Zimmer J, Bryl E, Goronzy JJ, et al. Premature telomeric loss in rheumatoid arthritis is genetically determined and involves both myeloid and lymphoid cell lineages. Proc Natl Acad Sci U S A. 2003;100(23):13471-6.
13. La Cava A, Nelson JL, Ollier WE, MacGregor A, Keystone EC, Thorne JC, et al. Genetic bias in immune responses to a cassette shared by different microorganisms in patients with rheumatoid arthritis. The Journal of clinical investigation. 1997;100(3):658-63.
14. Holoshitz J, Ling S. Nitric oxide signaling triggered by the rheumatoid arthritis shared epitope: a new paradigm for MHC disease association? Ann N Y Acad Sci. 2007;1110:73-83.
15. Yarwood A, Han B, Raychaudhuri S, Bowes J, Lunt M, Pappas DA, et al. A weighted genetic risk score using all known susceptibility variants to estimate rheumatoid arthritis risk. Ann Rheum Dis. 2015;74(1):170-6.
16. Eyre S, Bowes J, Diogo D, Lee A, Barton A, Martin P, et al. High-density genetic mapping identifies new susceptibility loci for rheumatoid arthritis. Nat Genet. 2012;44(12):1336-40.

17. Stahl EA, Raychaudhuri S, Remmers EF, Xie G, Eyre S, Thomson BP, et al. Genome-wide association study meta-analysis identifies seven new rheumatoid arthritis risk loci. Nat Genet. 2010;42(6):508-14.
18. Okada Y, Terao C, Ikari K, Kochi Y, Ohmura K, Suzuki A, et al. Meta-analysis identifies nine new loci associated with rheumatoid arthritis in the Japanese population. Nat Genet. 2012;44(5):511-6.
19. Okada Y, Wu D, Trynka G, Raj T, Terao C, Ikari K, et al. Genetics of rheumatoid arthritis contributes to biology and drug discovery. Nature. 2014;506(7488):376-81.
20. Danila MI, Laufer VA, Reynolds RJ, Yan Q, Liu N, Gregersen PK, et al. Dense Genotyping of Immune-Related Regions Identifies Loci for Rheumatoid Arthritis Risk and Damage in African Americans. Mol Med. 2017;23:177-87.
21. Arnson Y, Shoenfeld Y, Amital H. Effects of tobacco smoke on immunity, inflammation and autoimmunity. Journal of autoimmunity. 2010;34(3):J258-65.
22. Alpizar-Rodriguez D, Pluchino N, Canny G, Gabay C, Finckh A. The role of female hormonal factors in the development of rheumatoid arthritis. Rheumatology (Oxford). 2017;56(8):1254-63.
23. Amini L, Kalhor M, Haghighi A, Seyedfatemi N, Hosseini F. Effect of oral contraceptive pills on rheumatoid arthritis disease activity in women: A randomized clinical trial. Med J Islam Repub Iran. 2018;32:61.
24. Mikuls TR, Thiele GM, Deane KD, Payne JB, O'Dell JR, Yu F, et al. Porphyromonas gingivalis and disease-related autoantibodies in individuals at increased risk of rheumatoid arthritis. Arthritis Rheum. 2012;64(11):3522-30.
25. Bo M, Jasemi S, Uras G, Erre GL, Passiu G, Sechi LA. Role of Infections in the Pathogenesis of Rheumatoid Arthritis: Focus on Mycobacteria. Microorganisms. 2020;8(10).
26. Anzilotti C, Merlini G, Pratesi F, Tommasi C, Chimenti D, Migliorini P. Antibodies to viral citrullinated peptide in rheumatoid arthritis. J Rheumatol. 2006;33(4):647-51.
27. Sverdrup B, Kallberg H, Bengtsson C, Lundberg I, Padyukov L, Alfredsson L, et al. Association between occupational exposure to mineral oil and rheumatoid arthritis: results from the Swedish EIRA case-control study. Arthritis Res Ther. 2005;7(6):R1296-303.
28. Deane KD, Demoruelle MK, Kelmenson LB, Kuhn KA, Norris JM, Holers VM. Genetic and environmental risk factors for rheumatoid arthritis. Best Pract Res Clin Rheumatol. 2017;31(1):3-18.
29. Yeoh N, Burton JP, Suppiah P, Reid G, Stebbings S. The role of the microbiome in rheumatic diseases. Curr Rheumatol Rep. 2013;15(3):314.
30. Gross J, Oubaya N, Eymard F, Hourdille A, Chevalier X, Guignard S. Stressful life events as a trigger for rheumatoid arthritis onset within a year: a case-control study. Scand J Rheumatol. 2017;46(6):507-8.
31. Klareskog L, Stolt P, Lundberg K, Kallberg H, Bengtsson C, Grunewald J, et al. A new model for an etiology of rheumatoid arthritis: smoking may trigger HLA-DR (shared epitope)-restricted immune reactions to autoantigens modified by citrullination. Arthritis Rheum. 2006;54(1):38-46.
32. Anderton SM. Post-translational modifications of self antigens: implications for autoimmunity. Curr Opin Immunol. 2004;16(6):753-8.
33. Eggleton P, Nissim A, Ryan BJ, Whiteman M, Winyard PG. Detection and isolation of human serum autoantibodies that recognize oxidatively modified autoantigens. Free Radic Biol Med. 2013;57:79-91.

34. Burska AN, Hunt L, Boissinot M, Strollo R, Ryan BJ, Vital E, et al. Autoantibodies to posttranslational modifications in rheumatoid arthritis. Mediators Inflamm. 2014;2014:492873.
35. Vincent C, Nogueira L, Clavel C, Sebbag M, Serre G. Autoantibodies to citrullinated proteins: ACPA. Autoimmunity. 2005;38(1):17-24.
36. Vincent C, Serre G, Lapeyre F, Fournie B, Ayrolles C, Fournie A, et al. High diagnostic value in rheumatoid arthritis of antibodies to the stratum corneum of rat oesophagus epithelium, so-called 'antikeratin antibodies'. Ann Rheum Dis. 1989;48(9):712-22.
37. Scinocca M, Bell DA, Racape M, Joseph R, Shaw G, McCormick JK, et al. Antihomocitrullinated fibrinogen antibodies are specific to rheumatoid arthritis and frequently bind citrullinated proteins/peptides. J Rheumatol. 2014;41(2):270-9.
38. Shi J, van de Stadt LA, Levarht EW, Huizinga TW, Hamann D, van Schaardenburg D, et al. Anti-carbamylated protein (anti-CarP) antibodies precede the onset of rheumatoid arthritis. Ann Rheum Dis. 2014;73(4):780-3.
39. Holers VM. Autoimmunity to citrullinated proteins and the initiation of rheumatoid arthritis. Curr Opin Immunol. 2013;25(6):728-35.
40. Aletaha D, Alasti F, Smolen JS. Rheumatoid factor, not antibodies against citrullinated proteins, is associated with baseline disease activity in rheumatoid arthritis clinical trials. Arthritis Res Ther. 2015;17:229.
41. Sokolove J, Johnson DS, Lahey LJ, Wagner CA, Cheng D, Thiele GM, et al. Rheumatoid factor as a potentiator of anti-citrullinated protein antibody-mediated inflammation in rheumatoid arthritis. Arthritis & rheumatology. 2014;66(4):813-21.
42. Nielen MM, van Schaardenburg D, Reesink HW, van de Stadt RJ, van der Horst-Bruinsma IE, de Koning MH, et al. Specific autoantibodies precede the symptoms of rheumatoid arthritis: a study of serial measurements in blood donors. Arthritis Rheum. 2004;50(2):380-6.
43. Strollo R, Ponchel F, Malmstrom V, Rizzo P, Bombardieri M, Wenham CY, et al. Autoantibodies to posttranslationally modified type II collagen as potential biomarkers for rheumatoid arthritis. Arthritis Rheum. 2013;65(7):1702-12.
44. Deane KD, O'Donnell CI, Hueber W, Majka DS, Lazar AA, Derber LA, et al. The number of elevated cytokines and chemokines in preclinical seropositive rheumatoid arthritis predicts time to diagnosis in an age-dependent manner. Arthritis Rheum. 2010;62(11):3161-72.
45. Isaacs JD. The changing face of rheumatoid arthritis: sustained remission for all? Nat Rev Immunol. 2010;10(8):605-11.
46. Gerlag DM, Raza K, van Baarsen LG, Brouwer E, Buckley CD, Burmester GR. EULAR recommendations for terminology and research in individuals at risk of rheumatoid arthritis: report from the Study Group for Risk Factors for Rheumatoid Arthritis. Ann Rheum Dis. 2012;71.
47. Mankia K, Emery P. A new window of opportunity in rheumatoid arthritis: targeting at-risk individuals. Curr Opin Rheumatol. 2016;28(3):260-6.
48. Hunt L, Hensor EM, Nam J, Burska AN, Parmar R, Emery P, et al. T cell subsets: an immunological biomarker to predict progression to clinical arthritis in ACPA-positive individuals. Ann Rheum Dis. 2016;75(10):1884-9.
49. Catrina AI, Ytterberg AJ, Reynisdottir G, Malmstrom V, Klareskog L. Lungs, joints and immunity against citrullinated proteins in rheumatoid arthritis. Nature reviews Rheumatology. 2014;10(11):645-53.

50. Harre U, Georgess D, Bang H, Bozec A, Axmann R, Ossipova E, et al. Induction of osteoclastogenesis and bone loss by human autoantibodies against citrullinated vimentin. The Journal of clinical investigation. 2012;122(5):1791-802.
51. Bartok B, Firestein GS. Fibroblast-like synoviocytes: key effector cells in rheumatoid arthritis. Immunol Rev. 2010;233(1):233-55.
52. Udagawa N, Kotake S, Kamatani N, Takahashi N, Suda T. The molecular mechanism of osteoclastogenesis in rheumatoid arthritis. Arthritis Res. 2002;4(5):281-9.
53. Elshabrawy HA, Chen Z, Volin MV, Ravella S, Virupannavar S, Shahrara S. The pathogenic role of angiogenesis in rheumatoid arthritis. Angiogenesis. 2015;18(4):433-48.
54. Smolen JS, Aletaha D, Barton A, Burmester GR, Emery P, Firestein GS, et al. Rheumatoid arthritis. Nat Rev Dis Primers. 2018;4:18001.
55. Firestein GS. Pathogenesis of Rheumatoid Arthritis: The Intersection of Genetics and Epigenetics. Trans Am Clin Climatol Assoc. 2018;129:171-82.
56. Firestein GS. The immunopathogenesis of rheumatoid arthritis. Current Opinion in Rheumatology. 1991;3(3):398-406.
57. Siouti E, Andreakos E. The many facets of macrophages in rheumatoid arthritis. Biochem Pharmacol. 2019;165:152-69.
58. Roosnek E, Lanzavecchia A. Efficient and selective presentation of antigen-antibody complexes by rheumatoid factor B cells. J Exp Med. 1991;173(2):487-9.
59. Kinne RW, Brauer R, Stuhlmuller B, Palombo-Kinne E, Burmester GR. Macrophages in rheumatoid arthritis. Arthritis Res. 2000;2(3):189-202.
60. Smith MD, Kraan MC, Slavotinek J, Au V, Weedon H, Parker A, et al. Treatment-induced remission in rheumatoid arthritis patients is characterized by a reduction in macrophage content of synovial biopsies. Rheumatology. 2001;40(4):367-74.
61. Ohrndorf S, Glimm A-M, Burmester G, Backhaus M. Musculoskeletal ultrasound scoring systems: assessing disease activity and therapeutic response in rheumatoid arthritis. International Journal of Clinical Rheumatology. 2011;6:57-65.
62. Bugatti S, Vitolo B, Caporali R, Montecucco C, Manzo A. B cells in rheumatoid arthritis: from pathogenic players to disease biomarkers. Biomed Res Int. 2014;2014:681678.
63. Silverman GJ, Carson DA. Roles of B cells in rheumatoid arthritis. Arthritis Res Ther. 2003;5 Suppl 4:S1-6.
64. Trouw LA, Haisma EM, Levarht EW, van der Woude D, Ioan-Facsinay A, Daha MR, et al. Anti-cyclic citrullinated peptide antibodies from rheumatoid arthritis patients activate complement via both the classical and alternative pathways. Arthritis Rheum. 2009;60(7):1923-31.
65. Dusad A, Duryee MJ, Shaw AT, Klassen LW, Anderson DR, Wang D, et al. Induction of bone loss in DBA/1J mice immunized with citrullinated autologous mouse type II collagen in the absence of adjuvant. Immunol Res. 2014;58(1):51-60.
66. Edwards JC, Szczepanski L, Szechinski J, Filipowicz-Sosnowska A, Emery P, Close DR, et al. Efficacy of B-cell-targeted therapy with rituximab in patients with rheumatoid arthritis. The New England journal of medicine. 2004;350(25):2572-81.

67. Lee DSW, Rojas OL, Gommerman JL. B cell depletion therapies in autoimmune disease: advances and mechanistic insights. Nat Rev Drug Discov. 2020.
68. Pistoia V. Production of cytokines by human B cells in health and disease. Immunol Today. 1997;18(7):343-50.
69. Schultze JL, Michalak S, Lowne J, Wong A, Gilleece MH, Gribben JG, et al. Human Non-Germinal Center B Cell Interleukin (IL)-12 Production Is Primarily Regulated by T Cell Signals CD40 Ligand, Interferon γ, and IL-10: Role of B Cells in the Maintenance of T Cell Responses. Journal of Experimental Medicine. 1999;189(1):1-12.
70. Yeo L, Toellner KM, Salmon M, Filer A, Buckley CD, Raza K, et al. Cytokine mRNA profiling identifies B cells as a major source of RANKL in rheumatoid arthritis. Ann Rheum Dis. 2011;70(11):2022-8.
71. Lanchbury JS, Pitzalis C. Cellular Immune-Mechanisms in Rheumatoid-Arthritis and Other Inflammatory Arthritides. Current Opinion in Immunology. 1993;5(6):918-24.
72. Panayi GS, Lanchbury JS, Kingsley GH. The importance of the T cell in initiating and maintaining the chronic synovitis of rheumatoid arthritis. Arthritis Rheum. 1992;35(7):729-35.
73. Law SC, Street S, Yu CH, Capini C, Ramnoruth S, Nel HJ, et al. T-cell autoreactivity to citrullinated autoantigenic peptides in rheumatoid arthritis patients carrying HLA-DRB1 shared epitope alleles. Arthritis Res Ther. 2012;14(3):R118.
74. De Almeida DE, Ling S, Pi X, Hartmann-Scruggs AM, Pumpens P, Holoshitz J. Immune dysregulation by the rheumatoid arthritis shared epitope. J Immunol. 2010;185(3):1927-34.
75. Orozco G, Viatte S, Bowes J, Martin P, Wilson AG, Morgan AW, et al. Novel rheumatoid arthritis susceptibility locus at 22q12 identified in an extended UK genome-wide association study. Arthritis & rheumatology. 2014;66(1):24-30.
76. Plenge RM, Padyukov L, Remmers EF, Purcell S, Lee AT, Karlson EW, et al. Replication of putative candidate-gene associations with rheumatoid arthritis in > 4,000 samples from North America and Sweden: Association of susceptibility with PTPN22, CTLA4, and PADI4. Am J Hum Genet. 2005;77(6):1044-60.
77. Cobb JE, Plant D, Flynn E, Tadjeddine M, Dieude P, Cornelis F, et al. Identification of the Tyrosine-Protein Phosphatase Non-Receptor Type 2 as a Rheumatoid Arthritis Susceptibility Locus in Europeans. Plos One. 2013;8(6).
78. Kim K, Bang SY, Lee HS, Bae SC. Update on the genetic architecture of rheumatoid arthritis. Nature reviews Rheumatology. 2017;13(1):13-24.
79. Ha E, Bae SC, Kim K. Large-scale meta-analysis across East Asian and European populations updated genetic architecture and variant-driven biology of rheumatoid arthritis, identifying 11 novel susceptibility loci. Ann Rheum Dis. 2020.
80. Suzuki A, Terao C, Yamamoto K. Linking of genetic risk variants to disease-specific gene expression via multi-omics studies in rheumatoid arthritis. Semin Arthritis Rheum. 2019;49(3S):S49-S53.
81. Waase I, Kayser C, Carlson PJ, Goronzy JJ, Weyand CM. Oligoclonal T cell proliferation in patients with rheumatoid arthritis and their unaffected siblings. Arthritis Rheum. 1996;39(6):904-13.
82. Wagner UG, Koetz K, Weyand CM, Goronzy JJ. Perturbation of the T cell repertoire in rheumatoid arthritis. Proc Natl Acad Sci U S A. 1998;95(24):14447-52.

83. Castro-Sánchez P, Roda-Navarro P. Physiology and Pathology of Autoimmune Diseases: Role of CD4+ T cells in Rheumatoid Arthritis. Physiology and Pathology of Immunology: InTech; 2017.
84. Koetz K, Bryl E, Spickschen K, O'Fallon WM, Goronzy JJ, Weyand CM. T cell homeostasis in patients with rheumatoid arthritis. Proc Natl Acad Sci U S A. 2000;97(16):9203-8.
85. Symmons DPM, Farr M, Salmon M, Bacon PA. Lymphopenia in Rheumatoid Arthritis. Journal of the Royal Society of Medicine. 1989;82(8):462-3.
86. Goronzy JJ, Bartz-Bazzanella P, Hu W, Jendro MC, Walser-Kuntz DR, Weyand CM. Dominant clonotypes in the repertoire of peripheral CD4+ T cells in rheumatoid arthritis. The Journal of clinical investigation. 1994;94(5):2068-76.
87. Colmegna I, Diaz-Borjon A, Fujii H, Schaefer L, Goronzy JJ, Weyand CM. Defective proliferative capacity and accelerated telomeric loss of hematopoietic progenitor cells in rheumatoid arthritis. Arthritis Rheum. 2008;58(4):990-1000.
88. Martens PB, Goronzy JJ, Schaid D, Weyand CM. Expansion of unusual CD4+ T cells in severe rheumatoid arthritis. Arthritis Rheum. 1997;40(6):1106-14.
89. Namekawa T, Wagner UG, Goronzy JJ, Weyand CM. Functional subsets of CD4 T cells in rheumatoid synovitis. Arthritis Rheum. 1998;41(12):2108-16.
90. Schmidt D, Goronzy JJ, Weyand CM. CD4+ CD7- CD28- T cells are expanded in rheumatoid arthritis and are characterized by autoreactivity. The Journal of clinical investigation. 1996;97(9):2027-37.
91. Lawson CA, Brown AK, Bejarano V, Douglas SH, Burgoyne CH, Greenstein AS, et al. Early rheumatoid arthritis is associated with a deficit in the CD4+CD25high regulatory T cell population in peripheral blood. Rheumatology (Oxford). 2006;45(10):1210-7.
92. Ponchel F, Morgan AW, Bingham SJ, Quinn M, Buch M, Verburg RJ, et al. Dysregulated lymphocyte proliferation and differentiation in patients with rheumatoid arthritis. Blood. 2002;100(13):4550-6.
93. Ponchel F, Vital E, Kingsbury SR, El-Sherbiny YM. CD4+ T-cell subsets in rheumatoid arthritis.[Report]. International Journal of Clinical Rheumatology 7(1). 2012:37-53.
94. Burgoyne CH, Field SL, Brown AK, Hensor EM, English A, Bingham SL, et al. Abnormal T cell differentiation persists in patients with rheumatoid arthritis in clinical remission and predicts relapse. Annals of the Rheumatic Diseases. 2008;67(6):750-7.
95. Gul HL, Eugenio G, Rabin T, Burska A, Parmar R, Wu J, et al. Defining remission in rheumatoid arthritis: does it matter to the patient? A comparison of multi-dimensional remission criteria and patient reported outcomes. Rheumatology (Oxford). 2020;59(3):613-21.
96. Kawashima M, Miossec P. Defect of Th1 immune response of whole blood cells from active patients with rheumatoid arthritis (RA). Arthritis Research & Therapy. 2003;5:S14-S.
97. Kawashima M, Miossec P. Effect of treatment of rheumatoid arthritis with infliximab on IFN gamma, IL4, T-bet, and GATA-3 expression: link with improvement of systemic inflammation and disease activity. Ann Rheum Dis. 2005;64(3):415-8.
98. Ponchel F, Brown AK, Field SL, Quinn M, Conaghan P, Emery P, et al. T-bet expression in rheumatoid arthritis patients with early, disease-modifying anti-

rheumatic drug naïve disease is low and correlates with low levels of IL-7 and T-cell dysfunctions. Arthritis Research & Therapy. 2005;7(Suppl 1):P18-P.
99. Churchman SM, El-Jawhari JJ, Burska AN, Parmar R, Goeb V, Conaghan PG, et al. Modulation of peripheral T-cell function by interleukin-7 in rheumatoid arthritis. Arthritis Res Ther. 2014;16(6):511.
100. van Roon JAG, Glaudemans CA, Bijlsma JWJ, Lafeber F. Differentiation of naive CD4(+) T cells towards T helper 2 cells is not impaired in rheumatoid arthritis patients. Arthritis Research & Therapy. 2003;5(5):R269-R76.
101. Benghiat FS, Charbonnier LM, Vokaer B, De Wilde V, Le Moine A. Interleukin 17-producing T helper cells in alloimmunity. Transplant Rev (Orlando). 2009;23(1):11-8.
102. Arroyo-Villa I, Bautista-Caro M-B, Balsa A, Aguado-Acín P, Nuno L, Bonilla-Hernán M-G, et al. Frequency of Th17 CD4+ T cells in early rheumatoid arthritis: a marker of anti-CCP seropositivity. PLoS One. 2012;7(8):e42189.
103. Shen H, Goodall JC, Hill Gaston J. Frequency and phenotype of peripheral blood Th17 cells in ankylosing spondylitis and rheumatoid arthritis. Arthritis & Rheumatism. 2009;60(6):1647-56.
104. Truchetet M-E, Mossalayi MD, Boniface K. IL-17 in the rheumatologist's line of sight. BioMed research international. 2013;2013.
105. Behrens F, Himsel A, Rehart S, Stanczyk J, Beutel B, Zimmermann SY, et al. Imbalance in distribution of functional autologous regulatory T cells in rheumatoid arthritis. Ann Rheum Dis. 2007;66(9):1151-6.
106. Lim HW, Lee J, Hillsamer P, Kim CH. Human Th17 cells share major trafficking receptors with both polarized effector T cells and FOXP3+ regulatory T cells. The Journal of Immunology. 2008;180(1):122-9.
107. Annunziato F, Cosmi L, Santarlasci V, Maggi L, Liotta F, Mazzinghi B, et al. Phenotypic and functional features of human Th17 cells. The Journal of experimental medicine. 2007;204(8):1849-61.
108. Chalan P, Kroesen B-J, van der Geest KS, Huitema MG, Abdulahad WH, Bijzet J, et al. Circulating CD4+ CD161+ T lymphocytes are increased in seropositive arthralgia patients but decreased in patients with newly diagnosed rheumatoid arthritis. PloS one. 2013;8(11):e79370.
109. Robert M, Miossec P. IL-17 in Rheumatoid Arthritis and Precision Medicine: From Synovitis Expression to Circulating Bioactive Levels. Front Med (Lausanne). 2018;5:364.
110. Ciccia F, Guggino G, Rizzo A, Manzo A, Vitolo B, La Manna MP, et al. Potential involvement of IL-9 and Th9 cells in the pathogenesis of rheumatoid arthritis. Rheumatology (Oxford). 2015;54(12):2264-72.
111. Chowdhury K, Kumar U, Das S, Chaudhuri J, Kumar P, Kanjilal M, et al. Synovial IL-9 facilitates neutrophil survival, function and differentiation of Th17 cells in rheumatoid arthritis. Arthritis Res Ther. 2018;20(1):18.
112. Akdis M, Palomares O, van de Veen W, van Splunter M, Akdis CA. TH17 and TH22 cells: a confusion of antimicrobial response with tissue inflammation versus protection. J Allergy Clin Immunol. 2012;129(6):1438-49; quiz50-1.
113. Ponchel F, Burska AN, Hunt L, Gul H, Rabin T, Parmar R, et al. T-cell subset abnormalities predict progression along the Inflammatory Arthritis disease continuum: implications for management. Sci Rep. 2020;10(1):3669.
114. Ponchel F, Goeb V, Parmar R, El-Sherbiny Y, Boissinot M, El Jawhari J, et al. An immunological biomarker to predict MTX response in early RA. Annals of the Rheumatic Diseases. 2014;73(11):2047-53.

115. Niu Q, Cai B, Huang ZC, Shi YY, Wang LL. Disturbed Th17/Treg balance in patients with rheumatoid arthritis. Rheumatol Int. 2012;32(9):2731-6.
116. Ma J, Zhu C, Ma B, Tian J, Baidoo SE, Mao C, et al. Increased frequency of circulating follicular helper T cells in patients with rheumatoid arthritis. Clin Dev Immunol. 2012;2012:827480.
117. Monaco C, Nanchahal J, Taylor P, Feldmann M. Anti-TNF therapy: past, present and future. Int Immunol. 2015;27(1):55-62.
118. Geiler J, Buch M Fau - McDermott MF, McDermott MF. Anti-TNF treatment in rheumatoid arthritis. Current Pharmaceutical Design. 2011;17(29):3141 - 54.
119. Tanaka Y, Martin Mola E. IL-6 targeting compared to TNF targeting in rheumatoid arthritis: studies of olokizumab, sarilumab and sirukumab. Annals of the Rheumatic Diseases. 2014;73(9):1595-7.
120. Taylor PC. Clinical efficacy of launched JAK inhibitors in rheumatoid arthritis. Rheumatology (Oxford). 2019;58(Suppl 1):i17-i26.
121. Nikfar S, Saiyarsarai P, Tigabu BM, Abdollahi M. Efficacy and safety of interleukin-1 antagonists in rheumatoid arthritis: a systematic review and meta-analysis. Rheumatol Int. 2018;38(8):1363-83.
122. Clark W, Jobanputra P, Barton P, Burls A. The clinical and cost-effectiveness of anakinra for the treatment of rheumatoid arthritis in adults: a systematic review and economic analysis. Health Technol Assess. 2004;8(18):iii-iv, ix-x, 1-105.
123. Patel DD, Lee DM, Kolbinger F, Antoni C. Effect of IL-17A blockade with secukinumab in autoimmune diseases. Ann Rheum Dis. 2013;72 Suppl 2:ii116-23.
124. Martin DA, Churchill M, Flores-Suarez L, Cardiel MH, Wallace D, Martin R, et al. A phase Ib multiple ascending dose study evaluating safety, pharmacokinetics, and early clinical response of brodalumab, a human anti-IL-17R antibody, in methotrexate-resistant rheumatoid arthritis. Arthritis Res Ther. 2013;15(5):R164.
125. Choy EH, Panayi GS. Cytokine pathways and joint inflammation in rheumatoid arthritis. The New England journal of medicine. 2001;344(12):907-16.
126. Harrington LE, Hatton RD, Mangan PR, Turner H, Murphy TL, Murphy KM, et al. Interleukin 17-producing CD4+ effector T cells develop via a lineage distinct from the T helper type 1 and 2 lineages. Nat Immunol. 2005;6(11):1123-32.
127. Langrish CL, Chen Y, Blumenschein WM, Mattson J, Basham B, Sedgwick JD, et al. IL-23 drives a pathogenic T cell population that induces autoimmune inflammation. J Exp Med. 2005;201(2):233-40.
128. Nurieva R, Yang XO, Martinez G, Zhang Y, Panopoulos AD, Ma L, et al. Essential autocrine regulation by IL-21 in the generation of inflammatory T cells. Nature. 2007;448(7152):480-3.
129. Chung Y, Yang X, Chang SH, Ma L, Tian Q, Dong C. Expression and regulation of IL-22 in the IL-17-producing CD4+ T lymphocytes. Cell Res. 2006;16(11):902-7.
130. Ouyang W, Kolls JK, Zheng Y. The biological functions of T helper 17 cell effector cytokines in inflammation. Immunity. 2008;28(4):454-67.
131. Shahrara S, Pickens SR, Dorfleutner A, Pope RM. IL-17 induces monocyte migration in rheumatoid arthritis. J Immunol. 2009;182(6):3884-91.

132. Sato K, Suematsu A, Okamoto K, Yamaguchi A, Morishita Y, Kadono Y, et al. Th17 functions as an osteoclastogenic helper T cell subset that links T cell activation and bone destruction. J Exp Med. 2006;203(12):2673-82.
133. Psarras A, Emery P, Vital EM. Type I interferon-mediated autoimmune diseases: pathogenesis, diagnosis and targeted therapy. Rheumatology (Oxford). 2017;56(10):1662-75.
134. Crow M. Type I interferon in organ-targeted autoimmune and inflammatory diseases. . Arthritis Res Ther. 2010;12:S5.
135. Roelofs MF, Wenink MH, Brentano F, Abdollahi-Roodsaz S, Oppers-Walgreen B, Barrera P, et al. Type I interferons might form the link between Toll-like receptor (TLR) 3/7 and TLR4-mediated synovial inflammation in rheumatoid arthritis (RA). . Annals of the rheumatic diseases 2009;68:1486-93.
136. Van Holten J, Smeets T, Blankert P, PP. T. Expression of interferon β in synovial tissue from patients with rheumatoid arthritis: comparison with patients with osteoarthritis and reactive arthritis. . Annals of the rheumatic diseases 2005;64:1780-2.
137. Pilling D, Akbar AN, Girdlestone J, Orteu CH, Borthwick NJ, Amft N, et al. Interferon-beta mediates stromal cell rescue of T cells from apoptosis. European journal of immunology. 1999;29(3):1041-50.
138. Harada S, Yamamura M, Okamoto H, Morita Y, Kawashima M, Aita T. Production of interleukin-7 and interleukin-15 by fibroblast-like synoviocytes from patients with rheumatoid arthritis. Arthritis Rheum. 1999;42:1508-16.
139. van Roon JAG, Verweij MC, Wenting-van Wijk M, Jacobs KMG, Bijlsma JWJ, Lafeber F. Increased intraarticular interleukin-7 in rheumatoid arthritis patients stimulates cell contact-dependent activation of CD4 + T cells and macrophages. Arthritis Rheum. 2005;52.
140. Churchman SM, Ponchel F. Interleukin-7 in rheumatoid arthritis. Rheumatology. 2008;47(6):753-9.
141. Wehr P, Purvis H, Law SC, Thomas R. Dendritic cells, T cells and their interaction in rheumatoid arthritis. Clinical & Experimental Immunology. 2019;196(1):12-27.
142. McInnes IB, Leung BP, Liew FY. Cell-cell interactions in synovitis. Interactions between T lymphocytes and synovial cells. Arthritis research. 2000;2(5):374-8.
143. Brennan F, Foey A. Cytokine regulation in RA synovial tissue: role of T cell/macrophage contact-dependent interactions. Arthritis Res. 2002;4 Suppl 3:S177-82.
144. Kotake S, Udagawa N, Hakoda M, Mogi M, Yano K, Tsuda E, et al. Activated human T cells directly induce osteoclastogenesis from human monocytes: possible role of T cells in bone destruction in rheumatoid arthritis patients. Arthritis Rheum. 2001;44(5):1003-12.
145. Sawai H, Park YW, He X, Goronzy JJ, Weyand CM. Fractalkine mediates T cell-dependent proliferation of synovial fibroblasts in rheumatoid arthritis. Arthritis Rheum. 2007;56(10):3215-25.
146. Mor F, Quintana FJ, Cohen IR. Angiogenesis-inflammation cross-talk: vascular endothelial growth factor is secreted by activated T cells and induces Th1 polarization. J Immunol. 2004;172(7):4618-23.
147. Rao DA. T Cells That Help B Cells in Chronically Inflamed Tissues. Front Immunol. 2018;9:1924.

148. Isaacs JD, Manna VK, Rapson N, Bulpitt KJ, Hazleman BL, Matteson EL, et al. CAMPATH-1H in rheumatoid arthritis--an intravenous dose-ranging study. Br J Rheumatol. 1996;35(3):231-40.
149. Tyndall A, Gratwohl A. Hemopoietic blood and marrow transplants in the treatment of severe autoimmune disease. Current Opinion in Hematology. 1997;4(6).
150. Isaacs JD, Greer S, Sharma S, Symmons D, Smith M, Johnston J, et al. Morbidity and mortality in rheumatoid arthritis patients with prolonged and profound therapy-induced lymphopenia. Arthritis Rheum. 2001;44(9):1998-2008.
151. Jendro MC, Ganten T, Matteson EL, Weyand CM, Goronzy JJ. Emergence of Oligoclonal T-Cell Populations Following Therapeutic T-Cell Depletion in Rheumatoid-Arthritis. Arthritis Rheum. 1995;38(9):1242-51.
152. Weinblatt ME, Coblyn JS, Fox DA, Fraser PA, Holdsworth DE, Glass DN, et al. Efficacy of low-dose methotrexate in rheumatoid arthritis. The New England journal of medicine. 1985;312(13):818-22.
153. Szanto E. Low-dose methotrexate treatment of rheumatoid arthritis; long-term observation of efficacy and safety. Clin Rheumatol. 1989;8(3):323-20.
154. American College of Rheumatology Subcommittee on Rheumatoid Arthritis G. Guidelines for the management of rheumatoid arthritis: 2002 Update. Arthritis Rheum. 2002;46(2):328-46.
155. Kennedy T, McCabe C, Struthers G, Sinclair H, Chakravaty K, Bax D, et al. BSR guidelines on standards of care for persons with rheumatoid arthritis. Rheumatology. 2005;44(4):553-6.
156. Verstappen SM, Jacobs JW, van der Veen MJ, Heurkens AH, Schenk Y, ter Borg EJ, et al. Intensive treatment with methotrexate in early rheumatoid arthritis: aiming for remission. Computer Assisted Management in Early Rheumatoid Arthritis (CAMERA, an open-label strategy trial). Ann Rheum Dis. 2007;66(11):1443-9.
157. van den Broek M, Lems WF, Allaart CF. BeSt practice: the success of early-targeted treatment in rheumatoid arthritis. Clinical and experimental rheumatology. 2012;30(4 Suppl 73):S35-8.
158. Quinn MA, Emery P. Window of opportunity in early rheumatoid arthritis: possibility of altering the disease process with early intervention. Clinical and experimental rheumatology. 2003;21(5 Suppl 31):S154-7.
159. Smolen JS, Landewe R, Bijlsma J, Burmester G, Chatzidionysiou K, Dougados M, et al. EULAR recommendations for the management of rheumatoid arthritis with synthetic and biological disease-modifying antirheumatic drugs: 2016 update. Ann Rheum Dis. 2017;76(6):960-77.
160. National Institute for Health and Clinical Excellence (NICE). Rheumatoid arthritis in adults: management (NG100)2018. Available from: https://www.nice.org.uk/guidance/ng100.
161. Smolen JS, Aletaha D, Bijlsma JWJ, Breedveld FC, Boumpas D, Burmester G, et al. Treating rheumatoid arthritis to target: recommendations of an international task force. Annals of the Rheumatic Diseases. 2010;69(4):631.
162. Mackie S, Vital E, Ponchel F, Emery P. Co-stimulatory Blockade as Therapy for Rheumatoid Arthritis. Curr Rheumatol Reports. 2005;7:400-6.
163. Mason U, Aldrich J Fau - Breedveld F, Breedveld F Fau - Davis CB, Davis Cb Fau - Elliott M, Elliott M Fau - Jackson M, Jackson M Fau - Jorgensen C, et al. CD4 coating, but not CD4 depletion, is a predictor of efficacy with primatized

monoclonal anti-CD4 treatment of active rheumatoid arthritis. (0315-162X (Print)).
164. Cope AP, Schulze-Koops H, Aringer M. The central role of T cells in rheumatoid arthritis. Clinical and experimental rheumatology. 2007;25(5 Suppl 46):S4-11.
165. Yoshida Y, Tanaka T. Interleukin 6 and rheumatoid arthritis. BioMed research international. 2014;2014.
166. Angelini J, Talotta R, Roncato R, Fornasier G, Barbiero G, Dal Cin L, et al. JAK-Inhibitors for the Treatment of Rheumatoid Arthritis: A Focus on the Present and an Outlook on the Future. Biomolecules. 2020;10(7).
167. Zhang F, Wei K, Slowikowski K, Fonseka CY, Rao DA, Kelly S, et al. Defining inflammatory cell states in rheumatoid arthritis joint synovial tissues by integrating single-cell transcriptomics and mass cytometry. Nat Immunol. 2019;20(7):928-42.
168. Fonseka CY, Rao DA, Raychaudhuri S. Leveraging blood and tissue CD4+ T cell heterogeneity at the single cell level to identify mechanisms of disease in rheumatoid arthritis. Current Opinion in Immunology. 2017;49:27-36.
169. Cai S, Ming B, Ye C, Lin S, Hu P, Tang J, et al. Similar Transition Processes in Synovial Fibroblasts from Rheumatoid Arthritis and Osteoarthritis: A Single-Cell Study. Journal of Immunology Research. 2019;2019:4080735.
170. Rantapaa-Dahlqvist S, de Jong BA, Berglin E, Hallmans G, Wadell G, Stenlund H, et al. Antibodies against cyclic citrullinated peptide and IgA rheumatoid factor predict the development of rheumatoid arthritis. Arthritis Rheum. 2003;48(10):2741-9.
171. Spasovski D. Evaluation, hallmark, clinical relevance and role of anti citrullin antibody, IgG and IgM rheumatoid factor with serological parameters of disease activity or both as overall sifted test in undifferentiated seronegative arthropathy with or without joint inflammation and bone structure differences. Boost, over helling or amplification to classification criteria to rheumathoid arthritis? JOJ uro & nephron. 2016;1(2): 555557.
172. Angelotti F, Parma A, Cafaro G, Capecchi R, Alunno A, Puxeddu I. One year in review 2017: pathogenesis of rheumatoid arthritis. Clinical and experimental rheumatology. 2017;35(3):368-78.
173. Ponchel F, Burska AN. Epigenetic Modifications: Are we Closer to Clinical Applicability? Journal of Pharmacogenomics & Pharmacoproteomics. 2016;07(02).
174. Karouzakis E, Gay RE, Michel BA, Gay S, Neidhart M. DNA hypomethylation in rheumatoid arthritis synovial fibroblasts. Arthritis Rheum. 2009;60(12):3613-22.
175. Kuchen S, Seemayer CA, Rethage J, von Knoch R, Kuenzler P, Beat AM, et al. The L1 retroelement-related p40 protein induces p38delta MAP kinase. Autoimmunity. 2004;37(1):57-65.
176. Neidhart M, Rethage J, Kuchen S, Kunzler P, Crowl RM, Billingham ME, et al. Retrotransposable L1 elements expressed in rheumatoid arthritis synovial tissue: association with genomic DNA hypomethylation and influence on gene expression. Arthritis Rheum. 2000;43(12):2634-47.
177. Dupont C, Armant DR, Brenner CA. Epigenetics: definition, mechanisms and clinical perspective. Semin Reprod Med. 2009;27(5):351-7.
178. Bird A. Perceptions of epigenetics. Nature. 2007;447(7143):396-8.

179. Reik W, Dean W, Walter J. Epigenetic reprogramming in mammalian development. Science. 2001;293(5532):1089-93.
180. Holliday R, Pugh JE. DNA modification mechanisms and gene activity during development. Science. 1975;187(4173):226-32.
181. Sandovici I. Establishment of Tissue-Specific Epigenetic States During Development. In: Naumova AK, Greenwood CMT, editors. Epigenetics and Complex Traits. New York, NY: Springer New York; 2013. p. 35-62.
182. Szutorisz H, Canzonetta C, Georgiou A, Chow C-M, Tora L, Dillon N. Formation of an Active Tissue-Specific Chromatin Domain Initiated by Epigenetic Marking at the Embryonic Stem Cell Stage. Molecular and Cellular Biology. 2005;25(5):1804-20.
183. Sado T, Fenner MH, Tan SS, Tam P, Shioda T, Li E. X inactivation in the mouse embryo deficient for Dnmt1: distinct effect of hypomethylation on imprinted and random X inactivation. Dev Biol. 2000;225(2):294-303.
184. Payer B, Lee JT, Namekawa SH. X-inactivation and X-reactivation: epigenetic hallmarks of mammalian reproduction and pluripotent stem cells. Hum Genet. 2011;130(2):265-80.
185. Ospelt C, Gay S, Klein K. Epigenetics in the pathogenesis of RA. Seminars in immunopathology. 2017;39(4):409-19.
186. Ospelt C. Epigenetic biomarkers in rheumatology - the future? Swiss Med Wkly. 2016;146:w14312.
187. Baba S, Yamada Y, Hatano Y, Miyazaki Y, Mori H, Shibata T, et al. Global DNA hypomethylation suppresses squamous carcinogenesis in the tongue and esophagus. Cancer science. 2009;100(7):1186-91.
188. Esteller M. Aberrant DNA methylation as a cancer-inducing mechanism. Annual review of pharmacology and toxicology. 2005;45:629-56.
189. Herman JG, Baylin SB. Gene silencing in cancer in association with promoter hypermethylation. The New England journal of medicine. 2003;349(21):2042-54.
190. Nemtsova MV, Zaletaev DV, Bure IV, Mikhaylenko DS, Kuznetsova EB, Alekseeva EA, et al. Epigenetic Changes in the Pathogenesis of Rheumatoid Arthritis. Frontiers in genetics. 2019;10:570.
191. Cooper G. The Cell: A Molecular Approach. 2nd edition. Sunderland (MA): Sinauer Associates; 2000.
192. Li B, Carey M, Workman JL. The role of chromatin during transcription. Cell. 2007;128(4):707-19.
193. Golbabapour S, Abdulla MA, Hajrezaei M. A concise review on epigenetic regulation: insight into molecular mechanisms. Int J Mol Sci. 2011;12(12):8661-94.
194. Venkatesh S, Workman JL. Histone exchange, chromatin structure and the regulation of transcription. Nat Rev Mol Cell Biol. 2015;16(3):178-89.
195. Phillips T. The Role of Methylation in Gene Expression. Nature Education 1(1):116. 2008.
196. Okano M, Bell DW, Haber DA, Li E. DNA methyltransferases Dnmt3a and Dnmt3b are essential for de novo methylation and mammalian development. Cell. 1999;99(3):247-57.
197. Moore LD, Le T, Fan G. DNA methylation and its basic function. Neuropsychopharmacology. 2013;38(1):23-38.

198. Leonhardt H, Page AW, Weier HU, Bestor TH. A targeting sequence directs DNA methyltransferase to sites of DNA replication in mammalian nuclei. Cell. 1992;71(5):865-73.
199. Bird AP. DNA methylation and the frequency of CpG in animal DNA. Nucleic Acids Res. 1980;8(7):1499-504.
200. Bird AP. CpG-rich islands and the function of DNA methylation. Nature. 1986;321(6067):209-13.
201. Grewal SI, Moazed D. Heterochromatin and epigenetic control of gene expression. Science. 2003;301(5634):798-802.
202. Haberland M, Montgomery RL, Olson EN. The many roles of histone deacetylases in development and physiology: implications for disease and therapy. Nature reviews Genetics. 2009;10(1):32-42.
203. Choy MK, Movassagh M, Goh HG, Bennett MR, Down TA, Foo RS. Genome-wide conserved consensus transcription factor binding motifs are hyper-methylated. BMC Genomics. 2010;11:519.
204. Aoyama T, Okamoto T, Nagayama S, Nishijo K, Ishibe T, Yasura K, et al. Methylation in the core-promoter region of the chondromodulin-I gene determines the cell-specific expression by regulating the binding of transcriptional activator Sp3. The Journal of biological chemistry. 2004;279(27):28789-97.
205. Defossez PA, Stancheva I. Biological functions of methyl-CpG-binding proteins. Prog Mol Biol Transl Sci. 2011;101:377-98.
206. Sarkar S, Abujamra AL, Loew JE, Forman LW, Perrine SP, Faller DV. Histone deacetylase inhibitors reverse CpG methylation by regulating DNMT1 through ERK signaling. Anticancer Res. 2011;31(9):2723-32.
207. Richardson B, Yung R. Role of DNA methylation in the regulation of cell function. The Journal of laboratory and clinical medicine. 1999;134(4):333-40.
208. Ehrlich M. DNA hypomethylation in cancer cells. Epigenomics. 2009;1(2):239-59.
209. Dehan P, Kustermans G, Guenin S, Horion J, Boniver J, Delvenne P. DNA methylation and cancer diagnosis: new methods and applications. Expert review of molecular diagnostics. 2009;9(7):651-7.
210. Ellis L, Atadja PW, Johnstone RW. Epigenetics in cancer: targeting chromatin modifications. Molecular cancer therapeutics. 2009;8(6):1409-20.
211. Coyle YM, Xie XJ, Lewis CM, Bu D, Milchgrub S, Euhus DM. Role of physical activity in modulating breast cancer risk as defined by APC and RASSF1A promoter hypermethylation in nonmalignant breast tissue. Cancer Epidemiol Biomarkers Prev. 2007;16(2):192-6.
212. Chen H, Ke Q, Kluz T, Yan Y, Costa M. Nickel ions increase histone H3 lysine 9 dimethylation and induce transgene silencing. Mol Cell Biol. 2006;26(10):3728-37.
213. Ren X, McHale CM, Skibola CF, Smith AH, Smith MT, Zhang L. An emerging role for epigenetic dysregulation in arsenic toxicity and carcinogenesis. Environ Health Perspect. 2011;119(1):11-9.
214. Ravegnini G, Sammarini G, Hrelia P, Angelini S. Key Genetic and Epigenetic Mechanisms in Chemical Carcinogenesis. Toxicol Sci. 2015;148(1):2-13.
215. Zheng SC, Widschwendter M, Teschendorff AE. Epigenetic drift, epigenetic clocks and cancer risk. Epigenomics. 2016;8(5):705-19.
216. Walsh CP, Xu GL. Cytosine methylation and DNA repair. Curr Top Microbiol Immunol. 2006;301:283-315.

217. Ehrlich M, Norris KF, Wang RY, Kuo KC, Gehrke CW. DNA cytosine methylation and heat-induced deamination. Biosci Rep. 1986;6(4):387-93.
218. Elsayed GM, Fahmi AEA, Shafik NF, Elshimy RAA, Abd Elhakeem HK, Attea SA. Study of DNA methyl transferase 3A mutation in acute myeloid leukemic patients. Egyptian Journal of Medical Human Genetics. 2018;19(4):315-9.
219. Han M, Jia L, Lv W, Wang L, Cui W. Epigenetic Enzyme Mutations: Role in Tumorigenesis and Molecular Inhibitors. Front Oncol. 2019;9:194.
220. Komori HK, Hart T, LaMere SA, Chew PV, Salomon DR. Defining CD4 T cell memory by the epigenetic landscape of CpG DNA methylation. J Immunol. 2015;194(4):1565-79.
221. Cui H, Cruz-Correa M, Giardiello FM, Hutcheon DF, Kafonek DR, Brandenburg S, et al. Loss of IGF2 imprinting: a potential marker of colorectal cancer risk. Science. 2003;299(5613):1753-5.
222. Feinberg AP, Ohlsson R, Henikoff S. The epigenetic progenitor origin of human cancer. Nature reviews Genetics. 2006;7(1):21-33.
223. Wang Z, Chang C, Lu Q. Epigenetics of CD4+ T cells in autoimmune diseases. Curr Opin Rheumatol. 2017;29(4):361-8.
224. Fu LH, Ma CL, Cong B, Li SJ, Chen HY, Zhang JG. Hypomethylation of proximal CpG motif of interleukin-10 promoter regulates its expression in human rheumatoid arthritis. Acta pharmacologica Sinica. 2011;32(11):1373-80.
225. Kim YI, Logan JW, Mason JB, Roubenoff R. DNA hypomethylation in inflammatory arthritis: reversal with methotrexate. The Journal of laboratory and clinical medicine. 1996;128(2):165-72.
226. Maciejewska-Rodrigues H, Karouzakis E, Striethołt S, Hemmatazad H, Neidhart M, Ospelt C, et al. Epigenetics and rheumatoid arthritis: the role of SENP1 in the regulation of MMP-1 expression. Journal of autoimmunity. 2010;35(1):15-22.
227. Ishida K, Kobayashi T, Ito S, Komatsu Y, Yokoyama T, Okada M, et al. Interleukin-6 gene promoter methylation in rheumatoid arthritis and chronic periodontitis. Journal of periodontology. 2012;83(7):917-25.
228. Nile CJ, Read RC, Akil M, Duff GW, Wilson AG. Methylation status of a single CpG site in the IL6 promoter is related to IL6 messenger RNA levels and rheumatoid arthritis. Arthritis Rheum. 2008;58(9):2686-93.
229. Lian X, Xiao R, Hu X, Kanekura T, Jiang H, Li Y, et al. DNA demethylation of CD40l in CD4+ T cells from women with systemic sclerosis: a possible explanation for female susceptibility. Arthritis Rheum. 2012;64(7):2338-45.
230. Liao J, Liang G, Xie S, Zhao H, Zuo X, Li F, et al. CD40L demethylation in CD4(+) T cells from women with rheumatoid arthritis. Clinical immunology. 2012;145(1):13-8.
231. Lu Q, Wu A, Tesmer L, Ray D, Yousif N, Richardson B. Demethylation of CD40LG on the inactive X in T cells from women with lupus. J Immunol. 2007;179(9):6352-8.
232. Cribbs AP, Kennedy A, Penn H, Read JE, Amjadi P, Green P, et al. Treg cell function in rheumatoid arthritis is compromised by ctla-4 promoter methylation resulting in a failure to activate the indoleamine 2,3-dioxygenase pathway. Arthritis & rheumatology. 2014;66(9):2344-54.
233. Liebling MR. Methylation of the CTLA-4 promoter and Treg cell dysfunction in rheumatoid arthritis: comment on the article by Cribbs et al. Arthritis & rheumatology. 2015;67(5):1406.

234. Cribbs AP, Kennedy A, Penn H, Amjadi P, Green P, Read JE, et al. Methotrexate Restores Regulatory T Cell Function Through Demethylation of the FoxP3 Upstream Enhancer in Patients With Rheumatoid Arthritis. Arthritis & rheumatology. 2015;67(5):1182-92.
235. de Andres MC, Perez-Pampin E, Calaza M, Santaclara FJ, Ortea I, Gomez-Reino JJ, et al. Assessment of global DNA methylation in peripheral blood cell subpopulations of early rheumatoid arthritis before and after methotrexate. Arthritis Res Ther. 2015;17:233.
236. Guo S, Zhu Q, Jiang T, Wang R, Shen Y, Zhu X, et al. Genome-wide DNA methylation patterns in CD4+ T cells from Chinese Han patients with rheumatoid arthritis. Mod Rheumatol. 2017;27(3):441-7.
237. Glossop JR, Emes RD, Nixon NB, Haworth KE, Packham JC, Dawes PT, et al. Genome-wide DNA methylation profiling in rheumatoid arthritis identifies disease-associated methylation changes that are distinct to individual T- and B-lymphocyte populations. Epigenetics. 2014;9(9):1228-37.
238. Karouzakis E, Raza K, Kolling C, Buckley CD, Gay S, Filer A, et al. Analysis of early changes in DNA methylation in synovial fibroblasts of RA patients before diagnosis. Sci Rep. 2018;8(1):7370.
239. Clark A, Naamane N, Nair N, Anderson A, Skelton A, Diboll J, et al. P124 Altered CD4+ T cell DNA methylation in early rheumatoid arthritis. Annals of the Rheumatic Diseases. 2018;77(Suppl 1):A67-A.
240. Glossop JR, Emes RD, Nixon NB, Packham JC, Fryer AA, Mattey DL, et al. Genome-wide profiling in treatment-naive early rheumatoid arthritis reveals DNA methylome changes in T and B lymphocytes. Epigenomics. 2016;8(2):209-24.
241. Cortessis VK, Thomas DC, Levine AJ, Breton CV, Mack TM, Siegmund KD, et al. Environmental epigenetics: prospects for studying epigenetic mediation of exposure-response relationships. Hum Genet. 2012;131(10):1565-89.
242. Huang FY, Chan AO, Rashid A, Wong DK, Seto WK, Cho CH, et al. Interleukin-1beta increases the risk of gastric cancer through induction of aberrant DNA methylation in a mouse model. Oncol Lett. 2016;11(4):2919-24.
243. Gasche JA, Hoffmann J, Boland CR, Goel A. Interleukin-6 promotes tumorigenesis by altering DNA methylation in oral cancer cells. International journal of cancer. 2011;129(5):1053-63.
244. Hodge DR, Xiao W, Clausen PA, Heidecker G, Szyf M, Farrar WL. Interleukin-6 regulation of the human DNA methyltransferase (HDNMT) gene in human erythroleukemia cells. The Journal of biological chemistry. 2001;276(43):39508-11.
245. Hartnett L, Egan LJ. Inflammation, DNA methylation and colitis-associated cancer. Carcinogenesis. 2012;33(4):723-31.
246. Nakano K, Boyle DL, Firestein GS. Regulation of DNA methylation in rheumatoid arthritis synoviocytes. J Immunol. 2013;190(3):1297-303.
247. Bock C. Analysing and interpreting DNA methylation data. Nature reviews Genetics. 2012;13(10):705-19.
248. Kurdyukov S, Bullock M. DNA Methylation Analysis: Choosing the Right Method. Biology. 2016;5(1).
249. Dirks RA, Stunnenberg HG, Marks H. Genome-wide epigenomic profiling for biomarker discovery. Clinical epigenetics. 2016;8:122.

250. Bibikova M, Barnes B, Tsan C, Ho V, Klotzle B, Le JM, et al. High density DNA methylation array with single CpG site resolution. Genomics. 2011;98(4):288-95.
251. Dedeurwaerder S, Defrance M, Calonne E, Denis H, Sotiriou C, Fuks F. Evaluation of the Infinium Methylation 450K technology. Epigenomics. 2011;3(6):771-84.
252. Illumina. Infinium® HumanMethylation450 BeadChip [Available from: https://www.illumina.com/products/methylation_450_beadchip_kits.html.
253. Sandoval J, Heyn H, Moran S, Serra-Musach J, Pujana MA, Bibikova M, et al. Validation of a DNA methylation microarray for 450,000 CpG sites in the human genome. Epigenetics. 2011;6(6):692-702.
254. Gentleman RC, Carey VJ, Bates DM, Bolstad B, Dettling M, Dudoit S, et al. Bioconductor: open software development for computational biology and bioinformatics. Genome Biol. 2004;5(10):R80.
255. Huber W, Carey VJ, Gentleman R, Anders S, Carlson M, Carvalho BS, et al. Orchestrating high-throughput genomic analysis with Bioconductor. Nat Methods. 2015;12(2):115-21.
256. Nyren P, Pettersson B, Uhlen M. Solid phase DNA minisequencing by an enzymatic luminometric inorganic pyrophosphate detection assay. Anal Biochem. 1993;208(1):171-5.
257. Diggle MA, Clarke SC. Pyrosequencing: sequence typing at the speed of light. Mol Biotechnol. 2004;28(2):129-37.
258. Herman JG, Graff JR, Myohanen S, Nelkin BD, Baylin SB. Methylation-specific PCR: a novel PCR assay for methylation status of CpG islands. Proc Natl Acad Sci U S A. 1996;93(18):9821-6.
259. Eads CA, Danenberg KD, Kawakami K, Saltz LB, Blake C, Shibata D, et al. MethyLight: a high-throughput assay to measure DNA methylation. Nucleic Acids Res. 2000;28(8):E32.
260. Liloglou T, Nikolaidis G. Quantitative Methylation Specific PCR (qMSP). Bio-protocol. 2013;3(16):e871.
261. Wojdacz TK, Dobrovic A. Methylation-sensitive high resolution melting (MS-HRM): a new approach for sensitive and high-throughput assessment of methylation. Nucleic Acids Res. 2007;35(6):e41.
262. R Core Team. R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing; 2018.
263. Wickham H, François R, Henry L, Müller K. dplyr: A Grammar of Data Manipulation. R package version 0.8.5. 2020.
264. Zhang Z. Reshaping and aggregating data: an introduction to reshape package. Ann Transl Med. 2016;4(4):78.
265. Jr FEH. rms: Regression Modeling Strategies. R package version 5.1-4. 2019.
266. Turner SD. qqman: an R package for visualizing GWAS results using Q-Q and manhattan plots. bioRxiv. 2014.
267. Gregory RW, Ben B, Lodewijk B, Robert G, Wolfgang HAL, Thomas L, et al. gplots: Various R Programming Tools for Plotting Data. R package version 301. 2016.
268. Wickham H. ggplot2: Elegant Graphics for Data Analysis. . Springer-Verlag New York. 2016.
269. Aryee MJ, Jaffe AE, Corrada-Bravo H, Ladd-Acosta C, Feinberg AP, Hansen KD, et al. Minfi: a flexible and comprehensive Bioconductor package for

the analysis of Infinium DNA methylation microarrays. Bioinformatics. 2014;30(10):1363-9.
270. Tim T, Jr. FDb.InfiniumMethylation.hg19: Annotation package for Illumina Infinium DNA methylation probes. R package version 220. 2014.
271. Gentleman R, Carey V, Huber W, Hahne F. genefilter: genefilter: methods for filtering genes from high-throughput experiments. R package version 1581. 2017.
272. Jaffe AE, Murakami P, Lee H, Leek JT, Fallin MD, Feinberg AP, et al. Bump hunting to identify differentially methylated regions in epigenetic epidemiology studies. International journal of epidemiology. 2012;41(1):200-9.
273. Peters TJ, Buckley MJ, Statham AL, Pidsley R, Samaras K, V Lord R, et al. De novo identification of differentially methylated regions in the human genome. Epigenetics & Chromatin. 2015;8(1):6.
274. Hulsen T, de Vlieg J, Alkema W. BioVenn – a web application for the comparison and visualization of biological lists using area-proportional Venn diagrams. BMC Genomics. 2008;9(1):488.
275. Thomas PD, Campbell MJ, Kejariwal A, Mi H, Karlak B, Daverman R, et al. PANTHER: a library of protein families and subfamilies indexed by function. Genome Res. 2003;13(9):2129-41.
276. Doncheva NT, Morris JH, Gorodkin J, Jensen LJ. Cytoscape StringApp: Network Analysis and Visualization of Proteomics Data. J Proteome Res. 2019;18(2):623-32.
277. Chen YA, Lemire M, Choufani S, Butcher DT, Grafodatskaya D, Zanke BW, et al. Discovery of cross-reactive probes and polymorphic CpGs in the Illumina Infinium HumanMethylation450 microarray. Epigenetics. 2013;8(2):203-9.
278. Du P, Zhang X, Huang CC, Jafari N, Kibbe WA, Hou L, et al. Comparison of Beta-value and M-value methods for quantifying methylation levels by microarray analysis. BMC Bioinformatics. 2010;11:587.
279. Szklarczyk D, Franceschini A, Wyder S, Forslund K, Heller D, Huerta-Cepas J, et al. STRING v10: protein–protein interaction networks, integrated over the tree of life. Nucleic Acids Research. 2015;43(Database issue):D447-D52.
280. Szklarczyk D, Morris JH, Cook H, Kuhn M, Wyder S, Simonovic M, et al. The STRING database in 2017: quality-controlled protein-protein association networks, made broadly accessible. Nucleic Acids Res. 2017;45(D1):D362-D8.
281. Li P, Demirci F, Mahalingam G, Demirci C, Nakano M, Meyers BC. An integrated workflow for DNA methylation analysis. J Genet Genomics. 2013;40(5):249-60.
282. Frommer M, McDonald LE, Millar DS, Collis CM, Watt F, Grigg GW, et al. A genomic sequencing protocol that yields a positive display of 5-methylcytosine residues in individual DNA strands. Proc Natl Acad Sci U S A. 1992;89(5):1827-31.
283. Louise W, Yanxia B, Heidi C, Theodore D. Enzymatic Methyl-seq: The Next Generation of Methylome Analysis [cited 2018. Available from: https://www.neb.com/tools-and-resources/feature-articles/enzymatic-methyl-seq-the-next-generation-of-methylome-analysis.
284. Pitaksalee R, Burska AN, Ajaib S, Rogers J, Parmar R, Mydlova K, et al. Differential CpG DNA methylation in peripheral naive CD4(+) T-cells in early rheumatoid arthritis patients. Clinical epigenetics. 2020;12(1):54.

285. Heid CA, Stevens J, Livak KJ, Williams PM. Real time quantitative PCR. Genome Res. 1996;6(10):986-94.
286. Trinh BN, Long TI, Laird PW. DNA methylation analysis by MethyLight technology. Methods. 2001;25(4):456-62.
287. Sestakova S, Salek C, Remesova H. DNA Methylation Validation Methods: a Coherent Review with Practical Comparison. Biol Proced Online. 2019;21:19.
288. AppliedBiosystems. Real-time PCR handbook2014. Available from: https://www.thermofisher.com/content/dam/LifeTech/global/Forms/PDF/real-time-pcr-handbook.pdf.
289. Navarro E, Serrano-Heras G, Castano MJ, Solera J. Real-time PCR detection chemistry. Clin Chim Acta. 2015;439:231-50.
290. Ponchel F, Toomes C, Bransfield K, Leong FT, Douglas SH, Field SL, et al. Real-time PCR based on SYBR-Green I fluorescence: an alternative to the TaqMan assay for a relative quantification of gene rearrangements, gene amplifications and micro gene deletions. BMC Biotechnol. 2003;3:18.
291. Ririe KM, Rasmussen RP, Wittwer CT. Product differentiation by analysis of DNA melting curves during the polymerase chain reaction. Anal Biochem. 1997;245(2):154-60.
292. Holland PM, Abramson RD, Watson R, Gelfand DH. Detection of specific polymerase chain reaction product by utilizing the 5'----3' exonuclease activity of Thermus aquaticus DNA polymerase. Proc Natl Acad Sci U S A. 1991;88(16):7276-80.
293. Ruiz-Villalba A, van Pelt-Verkuil E, Gunst QD, Ruijter JM, van den Hoff MJ. Amplification of nonspecific products in quantitative polymerase chain reactions (qPCR). Biomol Detect Quantif. 2017;14:7-18.
294. Rodriguez A, Rodriguez M, Cordoba JJ, Andrade MJ. Design of primers and probes for quantitative real-time PCR methods. Methods Mol Biol. 2015;1275:31-56.
295. Ray P, Manach Yannick L, Riou B, Houle Tim T, Warner David S. Statistical Evaluation of a Biomarker. Anesthesiology. 2010;112(4):1023-40.
296. Bossuyt PM. Clinical validity: defining biomarker performance. Scand J Clin Lab Invest Suppl. 2010;242:46-52.
297. Grund B, Sabin C. Analysis of biomarker data: logs, odds ratios, and receiver operating characteristic curves. Curr Opin HIV AIDS. 2010;5(6):473-9.
298. McHugh M. The odds ratio: calculation, usage and interpretation. Biochem Med. 2009;19(2):120-6.
299. Skapenko A, Wendler J, Lipsky PE, Kalden JR, Schulze-Koops H. Altered memory T cell differentiation in patients with early rheumatoid arthritis. J Immunol. 1999;163(1):491-9.
300. Nakano K, Whitaker JW, Boyle DL, Wang W, Firestein GS. DNA methylome signature in rheumatoid arthritis. Ann Rheum Dis. 2013;72(1):110-7.
301. de la Rica L, Urquiza JM, Gomez-Cabrero D, Islam ABMMK, Lopez-Bigas N, Tegner J, et al. Identification of novel markers in rheumatoid arthritis through integrated analysis of DNA methylation and microRNA expression. Journal of autoimmunity. 2013;41:6-16.
302. Pratt AG, Swan DC, Richardson S, Wilson G, Hilkens CM, Young DA, et al. A CD4 T cell gene signature for early rheumatoid arthritis implicates interleukin 6-mediated STAT3 signalling, particularly in anti-citrullinated peptide antibody-negative disease. Annals of the rheumatic diseases. 2012.

303. Gruden K, Hren M, Herman A, Blejec A, Albrecht T, Selbig J, et al. A "crossomics" study analysing variability of different components in peripheral blood of healthy caucasoid individuals. Plos One. 2012;7(1):e28761.
304. Straub RH, Paimela L, Peltomaa R, Scholmerich J, Leirisalo-Repo M. Inadequately low serum levels of steroid hormones in relation to interleukin-6 and tumor necrosis factor in untreated patients with early rheumatoid arthritis and reactive arthritis. Arthritis Rheum. 2002;46(3):654-62.
305. Baillet A, Gossec L, Paternotte S, Etcheto A, Combe B, Meyer O, et al. Evaluation of Serum Interleukin-6 Level as a Surrogate Marker of Synovial Inflammation and as a Factor of Structural Progression in Early Rheumatoid Arthritis: Results From a French National Multicenter Cohort. Arthrit Care Res. 2015;67(7):905-12.
306. Madhok R, Crilly A, Watson J, Capell HA. Serum Interleukin-6 Levels in Rheumatoid-Arthritis - Correlations with Clinical and Laboratory Indexes of Disease-Activity. Annals of the Rheumatic Diseases. 1993;52(3):232-4.
307. Burska AN, El-Jawhari JJ, Wu J, Wakefield RJ, Marzo-Ortega H, Conaghan PG, et al. Receptor activator of nuclear factor kappa-Beta ligand (RANKL) serum levels are associated with progression to seropositive/negative rheumatoid arthritis. Clinical and experimental rheumatology. 2020.
308. Churchman SM, Geiler J, Parmar R, Horner EA, Church LD, Emery P, et al. Multiplexing immunoassays for cytokine detection in the serum of patients with rheumatoid arthritis: lack of sensitivity and interference by rheumatoid factor. Clinical and experimental rheumatology. 2012;30(4):534-42.
309. Ntranos A, Casaccia P. Bromodomains: Translating the words of lysine acetylation into myelin injury and repair. Neurosci Lett. 2016;625:4-10.
310. Lang F, Shumilina E. Regulation of ion channels by the serum- and glucocorticoid-inducible kinase SGK1. FASEB J. 2013;27(1):3-12.
311. Baban B, Liu JY, Mozaffari MS. SGK-1 regulates inflammation and cell death in the ischemic-reperfused heart: pressure-related effects. Am J Hypertens. 2014;27(6):846-56.
312. Rhead B, Holingue C, Cole M, Shao X, Quach HL, Quach D, et al. Rheumatoid Arthritis Naive T Cells Share Hypermethylation Sites With Synoviocytes. Arthritis & rheumatology. 2017;69(3):550-9.
313. Tian J, Chau C, Hales TG, Kaufman DL. GABAA receptors mediate inhibition of T cell responses. Journal of Neuroimmunology. 1999;96(1):21-8.
314. Mendu SK, Bhandage A, Jin Z, Birnir B. Different Subtypes of GABA-A Receptors Are Expressed in Human, Mouse and Rat T Lymphocytes. Plos One. 2012;7(8).
315. Tian JD, Yong J, Dang H, Kaufman DL. Oral GABA treatment downregulates inflammatory responses in a mouse model of rheumatoid arthritis. Autoimmunity. 2011;44(6):465-70.
316. Tian JD, Lu YX, Zhang HW, Chau CH, Dang HN, Kaufman DL. gamma-aminobutyric acid inhibits T cell autoimmunity and the development of inflammatory responses in a mouse type 1 diabetes model. J Immunol. 2004;173(8):5298-304.
317. Garaud S, Le Dantec C, Jousse-Joulin S, Hanrotel-Saliou C, Saraux A, Mageed RA, et al. IL-6 modulates CD5 expression in B cells from patients with lupus by regulating DNA methylation. J Immunol. 2009;182(9):5623-32.
318. Dienz O, Rincon M. The effects of IL-6 on CD4 T cell responses. Clinical immunology. 2009;130(1):27-33.

319. Teague TK, Marrack P, Kappler JW, Vella AT. IL-6 rescues resting mouse T cells from apoptosis. The Journal of Immunology. 1997;158(12):5791-6.
320. Lotz M, Jirik F, Kabouridis P, Tsoukas C, Hirano T, Kishimoto T, et al. B cell stimulating factor 2/interleukin 6 is a costimulant for human thymocytes and T lymphocytes. Journal of Experimental Medicine. 1988;167(3):1253-8.
321. Rochman I, Paul WE, Ben-Sasson S. IL-6 increases primed cell expansion and survival. The Journal of Immunology. 2005;174(8):4761-7.
322. Ponchel F, Burska A, Raschke E, Olek S, Emery P. A8.11 Th17 cells as a diagnostic biomarker for rheumatoid arthritis (RA): Pilot data using an epigenetic QPCR assay. Annals of the Rheumatic Diseases. 2016;75(Suppl 1):A69-A.
323. Veldhoen M, Hocking RJ, Atkins CJ, Locksley RM, Stockinger B. TGFβ in the Context of an Inflammatory Cytokine Milieu Supports De Novo Differentiation of IL-17-Producing T Cells. Immunity. 2006;24(2):179-89.
324. Serada S, Fujimoto M, Mihara M, Koike N, Ohsugi Y, Nomura S, et al. IL-6 blockade inhibits the induction of myelin antigen-specific Th17 cells and Th1 cells in experimental autoimmune encephalomyelitis. Proc Natl Acad Sci U S A. 2008;105(26):9041-6.
325. Korn T, Bettelli E, Gao W, Awasthi A, Jager A, Strom TB, et al. IL-21 initiates an alternative pathway to induce proinflammatory T(H)17 cells. Nature. 2007;448(7152):484-7.
326. Sutton CE, Lalor SJ, Sweeney CM, Brereton CF, Lavelle EC, Mills KH. Interleukin-1 and IL-23 induce innate IL-17 production from gammadelta T cells, amplifying Th17 responses and autoimmunity. Immunity. 2009;31(2):331-41.
327. Van Roon J, Glaudemans K, Bijlsma J, Lafeber F. Interleukin 7 stimulates tumour necrosis factor α and Th1 cytokine production in joints of patients with rheumatoid arthritis. Annals of the rheumatic diseases. 2003;62(2):113-9.
328. Muskardin TLW, Niewold TB. Type I interferon in rheumatic diseases. Nature reviews Rheumatology. 2018;14(4):214-28.
329. Rodriguez-Carrio J, Alperi-Lopez M, Lopez P, Ballina-Garcia FJ, Suarez A. Heterogeneity of the Type I Interferon Signature in Rheumatoid Arthritis: A Potential Limitation for Its Use As a Clinical Biomarker. Front Immunol. 2017;8:2007.
330. Ronnblom L. The importance of the type I interferon system in autoimmunity. Clinical and experimental rheumatology. 2016;34(4 Suppl 98):21-4.
331. Lübbers J, Brink M, van de Stadt LA, Vosslamber S, Wesseling JG, van Schaardenburg D, et al. The type I IFN signature as a biomarker of preclinical rheumatoid arthritis. Annals of the rheumatic diseases. 2013;72(5):776-80.
332. Bilgic H, Ytterberg SR, Amin S, McNallan KT, Wilson JC, Koeuth T, et al. Interleukin-6 and type I interferon-regulated genes and chemokines mark disease activity in dermatomyositis. Arthritis Rheum. 2009;60(11):3436-46.
333. Zimmermann M, Arruda-Silva F, Bianchetto-Aguilera F, Finotti G, Calzetti F, Scapini P, et al. IFNalpha enhances the production of IL-6 by human neutrophils activated via TLR8. Sci Rep. 2016;6:19674.
334. Mehta AK, Gracias DT, Croft M. TNF activity and T cells. Cytokine. 2018;101:14-8.
335. Lee PP, Fitzpatrick DR, Beard C, Jessup HK, Lehar S, Makar KW, et al. A critical role for Dnmt1 and DNA methylation in T cell development, function, and survival. Immunity. 2001;15(5):763-74.

336. He S, Tong Q, Bishop DK, Zhang Y. Histone methyltransferase and histone methylation in inflammatory T-cell responses. Immunotherapy. 2013;5(9):989-1004.
337. Unutmaz D, Baldoni F, Abrignani S. Human naive T cells activated by cytokines differentiate into a split phenotype with functional features intermediate between naive and memory T cells. Int Immunol. 1995;7(9):1417-24.
338. Thiel S, Sommer U, Kortylewski M, Haan C, Behrmann I, Heinrich PC, et al. Termination of IL-6-induced STAT activation is independent of receptor internalization but requires de novo protein synthesis. Febs Lett. 2000;470(1):15-9.
339. Thiel S, Dahmen H, Martens A, Muller-Newen G, Schaper F, Heinrich PC, et al. Constitutive internalization and association with adaptor protein-2 of the interleukin-6 signal transducer gp130. Febs Lett. 1998;441(2):231-4.
340. Nair N, Wilson AG, Barton A. DNA methylation as a marker of response in rheumatoid arthritis. Pharmacogenomics. 2017;18(14):1323-32.
341. Weissenbach M, Clahsen T, Weber C, Spitzer D, Wirth D, Vestweber D, et al. Interleukin-6 is a direct mediator of T cell migration. European journal of immunology. 2004;34(10):2895-906.
342. Eto D, Lao C, DiToro D, Barnett B, Escobar TC, Kageyama R, et al. IL-21 and IL-6 are critical for different aspects of B cell immunity and redundantly induce optimal follicular helper CD4 T cell (Tfh) differentiation. Plos One. 2011;6(3):e17739.
343. Johnston RJ, Poholek AC, DiToro D, Yusuf I, Eto D, Barnett B, et al. Bcl6 and Blimp-1 Are Reciprocal and Antagonistic Regulators of T Follicular Helper Cell Differentiation. Science. 2009;325(5943):1006-10.
344. Nurieva RI, Chung Y, Hwang D, Yang XO, Kang HS, Ma L, et al. Generation of T follicular helper cells is mediated by interleukin-21 but independent of T helper 1, 2, or 17 cell lineages. Immunity. 2008;29(1):138-49.
345. Yu M, Cavero V, Lu Q, Li H. Follicular helper T cells in rheumatoid arthritis. Clinical rheumatology. 2015;34(9):1489-93.
346. Firestein GS, Zvaifler NJ. How important are T cells in chronic rheumatoid synovitis? II. T cell-independent mechanisms from beginning to end. Arthritis Rheum. 2002;46(2):298-308.
347. Unutmaz D, Abrignani S. Cytokines Can Activate Resting T-Lymphocytes. Chall Mod Med. 1994;8:49-52.
348. Emery P. The Roche Rheumatology Prize Lecture. The optimal management of early rheumatoid disease: the key to preventing disability. Br J Rheumatol. 1994;33(8):765-8.
349. Burgers LE, Raza K, van der Helm-van Mil AH. Window of opportunity in rheumatoid arthritis - definitions and supporting evidence: from old to new perspectives. RMD Open. 2019;5(1):e000870.
350. Nell VPK, Machold KP, Eberl G, Stamm TA, Uffmann M, Smolen JS. Benefit of very early referral and very early therapy with disease-modifying anti-rheumatic drugs in patients with early rheumatoid arthritis. Rheumatology. 2004;43(7):906-14.
351. van der Linden MP, le Cessie S, Raza K, van der Woude D, Knevel R, Huizinga TW, et al. Long-term impact of delay in assessment of patients with early arthritis. Arthritis Rheum. 2010;62(12):3537-46.
352. Kyburz D, Gabay C, Michel BA, Finckh A, physicians of S-R. The long-term impact of early treatment of rheumatoid arthritis on radiographic

progression: a population-based cohort study. Rheumatology (Oxford). 2011;50(6):1106-10.
353. Aletaha D, Neogi T, Silman AJ, Funovits J, Felson DT, Bingham CO, 3rd, et al. 2010 rheumatoid arthritis classification criteria: an American College of Rheumatology/European League Against Rheumatism collaborative initiative. Ann Rheum Dis. 2010;69(9):1580-8.
354. Lee DM, Schur PH. Clinical utility of the anti-CCP assay in patients with rheumatic diseases. Ann Rheum Dis. 2003;62(9):870-4.
355. Sproston NR, Ashworth JJ. Role of C-Reactive Protein at Sites of Inflammation and Infection. Frontiers in immunology. 2018;9:754-.
356. Chang PY, Yang CT, Cheng CH, Yu KH. Diagnostic performance of anti-cyclic citrullinated peptide and rheumatoid factor in patients with rheumatoid arthritis. Int J Rheum Dis. 2016;19(9):880-6.
357. Bas S, Perneger TV, Kunzle E, Vischer TL. Comparative study of different enzyme immunoassays for measurement of IgM and IgA rheumatoid factors. Ann Rheum Dis. 2002;61(6):505-10.
358. Bas S, Perneger TV, Seitz M, Tiercy JM, Roux-Lombard P, Guerne PA. Diagnostic tests for rheumatoid arthritis: comparison of anti-cyclic citrullinated peptide antibodies, anti-keratin antibodies and IgM rheumatoid factors. Rheumatology (Oxford). 2002;41(7):809-14.
359. Schellekens GA, Visser H, de Jong BA, van den Hoogen FH, Hazes JM, Breedveld FC, et al. The diagnostic properties of rheumatoid arthritis antibodies recognizing a cyclic citrullinated peptide. Arthritis Rheum. 2000;43(1):155-63.
360. Benjamin O, Bansal P, Goyal A, Lappin SL. Disease Modifying Anti-Rheumatic Drugs (DMARD). StatPearls. Treasure Island (FL)2020.
361. Combe B, Landewe R, Daien CI, Hua C, Aletaha D, Alvaro-Gracia JM, et al. 2016 update of the EULAR recommendations for the management of early arthritis. Ann Rheum Dis. 2017;76(6):948-59.
362. Nurmohamed MT, Dijkmans BA. Efficacy, tolerability and cost effectiveness of disease-modifying antirheumatic drugs and biologic agents in rheumatoid arthritis. Drugs. 2005;65(5):661-94.
363. García-Giménez JL, Seco-Cervera M, Tollefsbol TO, Romá-Mateo C, Peiró-Chova L, Lapunzina P, et al. Epigenetic biomarkers: Current strategies and future challenges for their use in the clinical laboratory. Crit Rev Clin Lab Sci. 2017;54(7-8):529-50.
364. Voyias PD, Patel A, Arasaradnam RP. Chapter 10 - Epigenetic Biomarkers of Disease. In: Tollefsbol TO, editor. Medical Epigenetics. Boston: Academic Press; 2016. p. 159-76.
365. García-Giménez JL, Ushijima T, Tollefsbol TO. Chapter 1 - Epigenetic Biomarkers: New Findings, Perspectives, and Future Directions in Diagnostics. In: García-Giménez JL, editor. Epigenetic Biomarkers and Diagnostics. Boston: Academic Press; 2016. p. 1-18.
366. Peedicayil J. Epigenetic biomarkers in psychiatric disorders. Br J Pharmacol. 2008;155(6):795-6.
367. Wu H, Liao J, Li Q, Yang M, Zhao M, Lu Q. Epigenetics as biomarkers in autoimmune diseases. Clinical immunology. 2018;196:34-9.
368. Locke WJ, Guanzon D, Ma C, Liew YJ, Duesing KR, Fung KYC, et al. DNA Methylation Cancer Biomarkers: Translation to the Clinic. Frontiers in genetics. 2019;10:1150.

369. Pan Y, Liu G, Zhou F, Su B, Li Y. DNA methylation profiles in cancer diagnosis and therapeutics. Clin Exp Med. 2018;18(1):1-14.
370. Chen Y, Li J, Yu X, Li S, Zhang X, Mo Z, et al. APC gene hypermethylation and prostate cancer: a systematic review and meta-analysis. Eur J Hum Genet. 2013;21(9):929-35.
371. Wu T, Giovannucci E, Welge J, Mallick P, Tang WY, Ho SM. Measurement of GSTP1 promoter methylation in body fluids may complement PSA screening: a meta-analysis. Br J Cancer. 2011;105(1):65-73.
372. Payne SR. From discovery to the clinic: the novel DNA methylation biomarker (m)SEPT9 for the detection of colorectal cancer in blood. Epigenomics. 2010;2(4):575-85.
373. Dietrich D, Jung M, Puetzer S, Leisse A, Holmes EE, Meller S, et al. Diagnostic and prognostic value of SHOX2 and SEPT9 DNA methylation and cytology in benign, paramalignant and malignant pleural effusions. Plos One. 2013;8(12):e84225.
374. Wick W, Weller M, van den Bent M, Sanson M, Weiler M, von Deimling A, et al. MGMT testing--the challenges for biomarker-based glioma treatment. Nat Rev Neurol. 2014;10(7):372-85.
375. Bock C. Epigenetic biomarker development. Epigenomics. 2009;1(1):99-110.
376. Bennett MR, Devarajan P. Chapter 1 - Characteristics of an Ideal Biomarker of Kidney Diseases. In: Edelstein CL, editor. Biomarkers of Kidney Disease. San Diego: Academic Press; 2011. p. 1-24.
377. Sehouli J, Loddenkemper C, Cornu T, Schwachula T, Hoffmuller U, Grutzkau A, et al. Epigenetic quantification of tumor-infiltrating T-lymphocytes. Epigenetics. 2011;6(2):236-46.
378. Burska AN, Thu A, Parmar R, Bzoma I, Samans B, Raschke E, et al. Quantifying circulating Th17 cells by qPCR: potential as diagnostic biomarker for rheumatoid arthritis. Rheumatology (Oxford). 2019;58(11):2015-24.
379. Precision-for-Medicine. Epiontis ID Immune Cell Monitoring. p. https://www.precisionformedicine.com/specialty-lab-services/immune-monitoring/immune-monitoring-by-epiontis-id/.
380. Zhu H, Wu LF, Mo XB, Lu X, Tang H, Zhu XW, et al. Rheumatoid arthritis-associated DNA methylation sites in peripheral blood mononuclear cells. Ann Rheum Dis. 2019;78(1):36-42.
381. Julia A, Absher D, Lopez-Lasanta M, Palau N, Pluma A, Waite Jones L, et al. Epigenome-wide association study of rheumatoid arthritis identifies differentially methylated loci in B cells. Hum Mol Genet. 2017;26(14):2803-11.
382. Biosystems A. Gene Expression Assay Performance Guaranteed With the TaqMan Assays QPCR Guarantee Program.
383. Kahn SL, Ronnett BM, Gravitt PE, Gustafson KS. Quantitative methylation-specific PCR for the detection of aberrant DNA methylation in liquid-based Pap tests. Cancer. 2008;114(1):57-64.
384. Husseiny MI, Kuroda A, Kaye AN, Nair I, Kandeel F, Ferreri K. Development of a quantitative methylation-specific polymerase chain reaction method for monitoring beta cell death in type 1 diabetes. Plos One. 2012;7(10):e47942.
385. Siebert S, Lyall DM, Mackay DF, Porter D, McInnes IB, Sattar N, et al. Characteristics of rheumatoid arthritis and its association with major comorbid

conditions: cross-sectional study of 502 649 UK Biobank participants. RMD Open. 2016;2(1):e000267.
386. Kallberg H, Padyukov L, Plenge RM, Ronnelid J, Gregersen PK, van der Helm-van Mil AH, et al. Gene-gene and gene-environment interactions involving HLA-DRB1, PTPN22, and smoking in two subsets of rheumatoid arthritis. Am J Hum Genet. 2007;80(5):867-75.
387. Xu B, Lin J. Characteristics and risk factors of rheumatoid arthritis in the United States: an NHANES analysis. PeerJ. 2017;5:e4035.
388. Harrell Jr FE, Lee KL, Mark DB. Multivariable prognostic models: Issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. Statistics in Medicine. 1996;15(4):361-87.
389. Bartlett J. Adjusting for optimism/overfitting in measures of predictive ability using bootstrapping 2014 [Available from: https://thestatsgeek.com/2014/10/04/adjusting-for-optimismoverfitting-in-measures-of-predictive-ability-using-bootstrapping/.
390. Bondarenko V. The Bootstrap Approach to Managing Model Uncertainty 2015 [Available from: https://rpubs.com/vadimus/bootstrap.
391. Nesher G, Moore TL, Dorner RW. In vitro effects of methotrexate on peripheral blood monocytes: modulation by folinic acid and S-adenosylmethionine. Ann Rheum Dis. 1991;50(9):637-41.
392. Wang YC, Chiang EP. Low-dose methotrexate inhibits methionine S-adenosyltransferase in vitro and in vivo. Mol Med. 2012;18:423-32.
393. Nesher G, Moore TL. The in vitro effects of methotrexate on peripheral blood mononuclear cells. Modulation by methyl donors and spermidine. Arthritis Rheum. 1990;33(7):954-9.
394. Oosterom N, Griffioen PH, den Hoed MAH, Pieters R, de Jonge R, Tissing WJE, et al. Global methylation in relation to methotrexate-induced oral mucositis in children with acute lymphoblastic leukemia. Plos One. 2018;13(7):e0199574.
395. Nair N, Plant D, Verstappen SM, Isaacs JD, Morgan AW, Hyrich KL, et al. Differential DNA methylation correlates with response to methotrexate in rheumatoid arthritis. Rheumatology (Oxford). 2019.
396. Gosselt HR, van Zelst BD, de Rotte M, Hazes JMW, de Jonge R, Heil SG. Higher baseline global leukocyte DNA methylation is associated with MTX non-response in early RA patients. Arthritis Res Ther. 2019;21(1):157.
397. Hazlewood GS, Barnabe C, Tomlinson G, Marshall D, Devoe D, Bombardier C. Methotrexate monotherapy and methotrexate combination therapy with traditional and biologic disease modifying antirheumatic drugs for rheumatoid arthritis: abridged Cochrane systematic review and network meta-analysis. BMJ. 2016;353:i1777.
398. Leoni C, Vincenzetti L, Emming S, Monticelli S. Epigenetics of T lymphocytes in health and disease. Swiss Med Wkly. 2015;145:w14191.
399. Coit P, Jeffries M, Altorok N, Dozmorov MG, Koelsch KA, Wren JD, et al. Genome-wide DNA methylation study suggests epigenetic accessibility and transcriptional poising of interferon-regulated genes in naive CD4+ T cells from lupus patients. Journal of autoimmunity. 2013;43:78-84.
400. Altorok N, Coit P, Hughes T, Koelsch KA, Stone DU, Rasmussen A, et al. Genome-wide DNA methylation patterns in naive CD4+ T cells from patients with primary Sjogren's syndrome. Arthritis & rheumatology. 2014;66(3):731-9.

401. Dong L, Wang X, Tan J, Li H, Qian W, Chen J, et al. Decreased expression of microRNA-21 correlates with the imbalance of Th17 and Treg cells in patients with rheumatoid arthritis. J Cell Mol Med. 2014;18(11):2213-24.
402. Ouboussad L, Hunt L, Hensor EMA, Nam JL, Barnes NA, Emery P, et al. Profiling microRNAs in individuals at risk of progression to rheumatoid arthritis. Arthritis Res Ther. 2017;19(1):288.
403. Ganapathi SK, Beggs AD, Hodgson SV, Kumar D. Expression and DNA methylation of TNF, IFNG and FOXP3 in colorectal cancer and their prognostic significance. Br J Cancer. 2014;111(8):1581-9.
404. Arroyo-Jousse V, Garcia-Diaz DF, Codner E, Perez-Bravo F. Epigenetics in type 1 diabetes: TNFa gene promoter methylation status in Chilean patients with type 1 diabetes mellitus. Br J Nutr. 2016;116(11):1861-8.
405. Campion J, Milagro FI, Goyenechea E, Martinez JA. TNF-alpha promoter methylation as a predictive biomarker for weight-loss response. Obesity (Silver Spring). 2009;17(6):1293-7.

List of supplementary data

(Excel file)

Data S1-S3 : DM-CpG-clusters and isolated-DM-CpG of naïve, memory CD4+T-cells and monocytes.

Data S4: final gene list add in string network analysis for 3 cells subsets

Data S5: Overlapping of DM gene list between 3 strategies; Scoring system, DMRcate, and Bumphunter.

Appendix

Appendix 1

Reagents, kits, buffer and equipment list.

Reagents and kits

Technique	Kit	Reagent	Company	Cat No
CD4+isolation	EasySep™ Human CD4+ T Cell	EasySep™ Human CD4+ T Cell Isolation Cocktail EasySep™ Dextran RapidSpheres™ 50103	Stemcell Technology	17952
Cell Isolation		LymphoprepTM	Alere Technologies AS	7801
		PBS	Sigma	P4417
		Trypan blue	Gibco™	15250061
		FAC Buffer *		
		Modified FAC Buffer (+2mM EDTA)*		
		Blocking Buffer *		
DNA isolation	QIAamp DNA Blood Mini Kit (250)	QIAamp Mini Spin Columns Collection Tubes (2 ml) Buffer AL Proteinase K Buffer AW1 Buffer AW2 Buffer AE	Qaigen	51106
Bisulfite conversion	EZ DNA Methylation-Gold Kit	M-dilution buffer M-dissolving buffer CT coversion reagent Desulphonation buffer Binding Elution Wash buffer	Zymo research ,Valencia, CA, US)	D5006
PCR	HotStarTaq DNA Polymerase	HotStarTaq DNA Polymerase 10x PCR Buffer 25 mM MgCl2	Qaigen	203205
		dNTP Mix, PCR Grade (200 ul)	Qaigen	201900
		Nuclease-Free Water (10 x 50 ml)	Qaigen	129114
Agarose electrophoresis		Agarose	Bioline	BIO-41025
		TBE BUFFER (10X)	Applichem Lifescience	A0972.1000
		50 bp DNA ladder	New England Biolabs	B7025S
		Hyper ladder 100bp	Bioline	BIO-33056
		Gel loadding dye	New England Biolabs	N3236S
		Ethidium Bromide (ETBr)	Sigma	E1385
Sequencing		illustra™ ExoProStar™	GE healthcare	US77705
	BigDye™ Terminator	ABI sequencing buffer Ready reaction mix	Applied Biosystems	4337454
		3M sodium acetate (pH 5.2)	Sigma	S7899
		Ethanol	Sigma	32221
		HiDi formamide	Applied Biosystems	4401457
Standard DNA		CpGmethylated Hela Genomic DNA	New England biolabs	N4007S
		Cells-to-CpG Methylated & Unmethylated gDNA Control 100 ng/ul g DNA	Applied Biosystems	4445552
		EpiTect PCR Control DNA Set (100)	Qaigen	59695
qPCR		Power SYBR™ Green PCR Master Mix	Applied Biosystems	4368706
		TaqMan™ Universal Master Mix II, no UNG	Applied Biosystems	4440047

Notes, Buffer recipes	
FAC Buffer *	500 mL (0.1% BSA in PBS (0.5g)+0.01% Sodium Azide +200 NaEDTA 0.5M
Modified FAC Buffer *	FAC Buffer + 2mM EDTA
Blocking Buffer *	Mouse serum(sigma-M5905)300 uL+ Human IgG(sigma-12511)100uL+FAC buffer1100

Equipment/Machine

Name	Company	Cat no.
Attune	Thermofisher	
Influx FACS	BD Biosciences	
ND1000 Spectrophotometer	NanoDrop Technologies	
ChemiDoc Imaging Systems	Bio-Rad Laboratories	
Techne™ TC-512 Gradient Thermal Cycler	Bibby Scientific	
3130xl Genetic Analyzer	Applied Biosystems	
Quant Studio5	Applied Biosystems	
MicroAmp optical 96-well reaction plate	Applied Biosystems	N8010560
MicroAmp optical adhesive flim	Applied Biosystems	4311971
EasySep™ Magnet	Stemcell Technology	18000

Appendix 2

Samples ethical permission

NHS

National Research Ethics Service

Leeds (West) Research Ethics Committee
Room 22
Floor CD, Block 40
King Edward Home
Leeds General Infirmary
Leeds
LS1 3EX

Telephone: 0113 3923181
Facsimile: 0113 3926799

15 January 2010

Professor Paul Emery
ARC Professor
Department of Rheumatology. Second Floor.
Chapel Allerton Hospital, Chapeltown Road
Leeds
LS7 4SA

Dear Professor Emery

Study Title:	**Inflammatory arthritis disease continuum longitudinal study**
REC reference number:	**09/H1307/98**
Protocol number:	**1**

Thank you for your letter of 11 January 2010, responding to the Committee's request for further information on the above research and submitting revised documentation.

The further information was considered in correspondence by a sub-committee of the REC A list of the sub-committee members is attached.

Confirmation of ethical opinion

On behalf of the Committee, I am pleased to confirm a favourable ethical opinion for the above research on the basis described in the application form, protocol and supporting documentation as revised, subject to the conditions specified below.

Ethical review of research sites

The favourable opinion applies to all NHS sites taking part in the study, subject to management permission being obtained from the NHS/HSC R&D office prior to the start of the study (see "Conditions of the favourable opinion" below).

Conditions of the favourable opinion

The favourable opinion is subject to the following conditions being met prior to the start of the study.

Management permission or approval must be obtained from each host organisation prior to the start of the study at the site concerned.

For NHS research sites only, management permission for research ("R&D approval") should be obtained from the relevant care organisation(s) in accordance with NHS research

Appendix 3

Supplementary method: Initial procedures for obtaining Genome-wide DNA methylation data: Samples procedure and data acquisition

Blood Samples were collected from 6 HC and 10 early patients (who meet the EULAR-2010 classification criteria for RA (1)), being naïve for DMARD, and having an active disease with at least 3 swollen joints and inflammation markers (CRP>10) (using a ficoll method). 3 cell subtypes, Naïve (CD45RA+/CD45RO-) and memory (CD45RA-/CD45RO+) CD4+T-cell and monocytes were purified by cell sorting following antibody staining using standards protocols: anti-CD4 (Clone RPA-T4, BD), anti-CD3 (Clone RPA-T8, BD), CD45RA (Clone MEM55, Serotec), CD45RO (Clone UCHL1, Serotec). Monocytes were purified based on the expression of CD14 detected by cell surface staining (CD14 Clone M5E2, BD). CD45RB/CD45RA/CD62L were used for selecting naïve and memory CD4+T-cells. Genomic DNA was extracted and bisulfide converted before a genome wide DNA methylation array was performed using an Illumina Infinium Human Methylation 450 Bead Chip. Raw data from the Illumina array were then forwarded to the research group.

Appendix 4

Flow cytometry antibodies staining detail

Technique	Target cell	Antibodies	Company	Cat no.
Cells sorting	CD4+T-cell	FITC Mouse Anti-Human CD3 Clone HIT3a (RUO)	BD	555339
		aPC-Cy™7 Mouse Anti-Human CD4 Clone SK3 (CE-IVD)	Biolegend	341115
	CD8+T-cell	FITC Mouse Anti-Human CD3 Clone HIT3a (RUO)	BD	555339
		Alexa Fluor® 700 Mouse Anti-Human CD8 Clone RPA-T8 (RUO)	BD	557945
	B-cell	APC Mouse Anti-Human CD19 Clone HIB19 (RUO)	BD	555415
	NK cell	PE Mouse Anti-Human CD56 Clone B159 (RUO)	BD	555516
	Monocyte	Pacific Blue™ Mouse Anti-Human CD14 Clone M5E2 (RUO)	BD	558121
Purification check	CD4+T-cell	FITC Mouse Anti-Human CD3 Clone HIT3a (RUO)	BD	555339
		V500 Mouse Anti-Human CD4 Clone RPA-T4 (RUO)	BD	560768

Appendix 5

Primers detail for Bisulfite sequencing

Gene	Position	F/R/Probe	sequcene	Product size
Bisulfite sequecing				
TNF	Chr 6: 31,543,118-31,543,391	F 5' to 3'	GAGTGTGAGGGGTATTTTTGATGTT	274
		R 5' to 3'	CTCTCCCTCTTAACTAATCCTCTA CTATCC	
Expected Product				
	GAGTGTGAGGGGTATTTTTGATGTT			
	TGTGTGTTTTTAATTTTTTAAATTTT **CG**TTTT**CGCG**ATGGAGAAGAAAT**CG**			
	AGATAGAAGGTGTAGGGTTTATTAT**CG**TTTTTTTTAGATGAGTTTATGGG			
	TTTTTTTATTAAGGAAGTTTTT**CG**TTGGTTGAATGATTTTTTTTT**CG**TTT			
	TTTTTT**CG**TTTTAGGGATATATAAAGGTAGTTGTTGGTATATTTAGTTAG			
	TAGA**CG**TTTTTTTAGTAAGGATAGTAGAGGATTAGTTAAGAGGGAGAG			
IFITM1	Chr 11: 315,655-315,886	F 5' to 3'	TTAGGTGTGTTGGATTTTAGTAGTTG	232
		R 5' to 3'	CCCGTCACATTTCAAAACTA	
		R 5' to 3'	CCTGTCACATTTCAAAACTA	
Expected Product				
	TTAGG TGTGTTGGAT TTTAGTAGTT			
	GTTTTTTTTA GTTTAGT**CG**G TATTTTTTTT GTTTTTTTGG GGTGGGGGTA			
	GTTAATGGTT **CG** AGGGTGGG G**CG**TT**CGCG**A GGGTTTGGGA GGGTAGTTTT			
	CGAT**CG**GGAG TTT**CGCG**GGA GTT**CG**GGAAG AGGGTTT**CG**T TAGGTGGAGA			
	TTTTTTTGGT GTAGTTAG**CG** AGGGTTT**CG**G GATTT**CG**TAG TTTTGAAATG TGA**CG**GG			

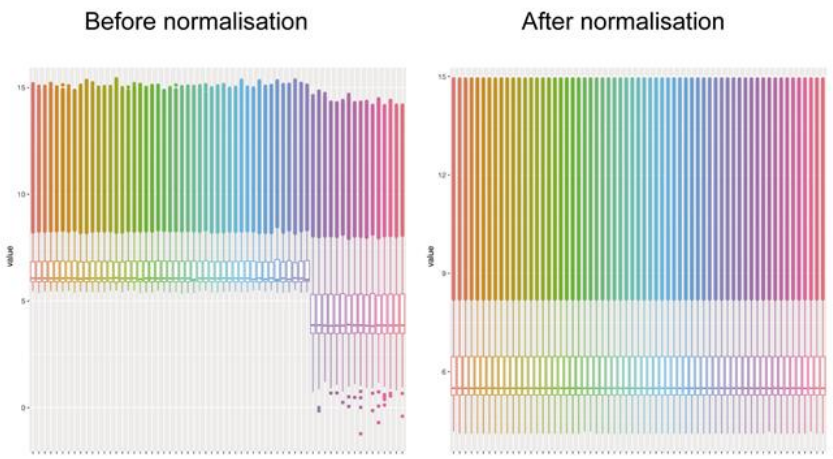
Primers and probes detail for qMSP assay

Gene	Position	F/R/Probe	sequcene	Product size
qMSP-SYBRgreen				
Target gene assay				
HDAC4_V1	Chr 2: 240,196,959-240,197,150	F 5' to 3'	**G**TTGGAGTTAGGT**TG**TTGGTAAAG**TG**	192
		R 5' to 3'	CC**CA**ACCAACAA**CA**CCTC**CA**A	
HDAC4_V2	Chr 2: 240,197,021-240,197,148	F 5' to 3'	GA**TG**TAAATTAAGAGGTAAGTAAATGA	128
		R 5' to 3'	**CA**ACCAACAA**CA**CCTC**CA**A	
HDAC4_V3	Chr 2: 240,196,858-240,196,984	F 5' to 3'	GGT**TG**GAAGTTTTGTGGGT**TG**A	127
		R 5' to 3'	**CA**CTTTAC**CA**ACAACCTAACTCCA	
PRTOR	Chr 17 : 78,786,209-78,786,379	F 5' to 3'	TTGTGAAGTTTGAATTGTTTTGAGT	171
		R 5' to 3'	CAA**CA**AAAACACTAAATAAATCA**CA**ACA	
MIR21	Chr 17: 57,915,689-57,915,816	F 5' to 3'	GTTATTTTAGGAGTATTTGGAGGTTTAA**TG**	128
		R 5' to 3'	AATATAATTTCAACCCAAAACTTCCCAA	
TNF_V1	Chr 6: 31,543,091-31,543,203	F 5' to 3'	TTT**TG**GAAT**TG**GAGTAGGGAGGA	113
		R 5' to 3'	CCTTCTATCT**CA**ATTTCTTCTCCAT**CACA**A	
TNF_V2	Chr 6: 31,544,670-31,544,842	F 5' to 3'	TTTGGAGATAATGTGAGAAGGATTT**GT**	173
Internal control assay				
GAPDH v1	Chr 12: 6,645,437-6,645,570	F 5' to 3'	TGATTGGGGGTGTTGGGTAGT	134
		R 5' to 3'	AATACAACATCTCCTTACCCCCAAA	
GAPDH v2	Chr 12: 6,645,384-6,645,511	F 5' to 3'	AGATTGTGGGTGGTAGTGTTT	128
		R 5' to 3'	CTTCCCTACCAAACTAACCTAACT	
ACTB V1	Chr 7: 5,565,747-5,565,939	F 5' to 3'	TTTAGTAGAGGGAAGGTAGGTTAGGTT	193
		R 5' to 3'	CCATAACTCACTACAACCTCAACTTT	
ACTB V2	Chr 7: 5,564,028-5,564,200	F 5' to 3'	TTGGTAAGAAGTAGGAGTTGTTGAAGTT	173
		F 5' to 3'	TTGGTAAGAAGTAGGAGTTGTTGAAGT	
ACTB V3	Chr 7: 5,565,650-5,565,765	F 5' to 3'	TGTTATGTATTAGGTGGTGTTTGGG	116
		R 5' to 3'	CTACCTTCCCTCTACTAAAAACT	
qMSP-Taq-Man				
Target gene assay				
TNF	Chr 6:31,543,091-31,543,211	F 5' to 3'	TTT**CG**GAAT**CG**GAGTAGGGAG	121
		R 5' to 3'	ACCCTACACCTTCTATCT**CG**ATTTCTT	
		Probe	T**CG**TTTT**CGCG**ATGGAG	
HDAC4	Chr 2:240,196,872-240,196,954	F 5' to 3'	TGGGT**CG**AAGTTATTTTAGGTTTTTAGT	83
		R 5' to 3'	AA**CG**ACTTACCAAAAACAACCTCAA	
IRF8 V2	Chr 16:85,979,046-85,979,116	F 5' to 3'	TGAAGTAGTAGTTT**CG**GTATTGGGTTT	71
		R 5' to 3'	ACCAACCCA**CG**CCAAAAA	
		Probe	TAGTGGAGAT**CG**GGAATGA	
IRF8 V4	Chr 16:85,979,046-85,979,126	F 5' to 3'	TGAAGTAGTAGTTT**CG**GTATTGGGTTT	81
		R 5' to 3'	CTA**CG**TCCTTACCAACCCA**CG**	
		Probe	TAGTGGAGAT**CG**GGAATGA	
Internal control assay				
GAPDH	Chr 12:6,645,449-6,645,570	F 5' to 3'	TTGGGTAGTTTTGGAGTTTTTAGTTG	122
		R 5' to 3'	AATACAACATCTCCTTACCCCCAA	
		Probe	AGTTAGGTTAGTTTGGTAGGGAA	
ACTB	Chr 7:5,571,788-5,571,861	F 5' to 3'	TGGTGATGGAGGAGGTTTAGTAAGT	74
		R 5' to 3'	TAACCACCACCCAACACACAAT	
		Probe	TGGATTGTGAATTTGTGTTTG	

Appendix 6

Supplementary method: Analysis of publicly available gene expression data

Gene expression data were obtained from ArrayExpress (E-GEOD-20098, E-GEOD-26163) (2, 3). Sample were CD4+T-cells, purified from peripheral blood of 47 early, drug naïve RA patients and 16 HC. These datasets were both generated using Illumina HumanWG-6 v3.0 expression beadchips and BeadStudio version 3.3.7 software (Illumina, San Diego, California, USA). Pre-processing of raw data was performed using the *Lumi* package (4, 5) for data input, quality assessment, and normalisation (robust spline) on $\log_2$-transfromed data. The *normalizeBetweenArrays* function within the Limma Package was used to achieve consistency between the two arrays (6) (Figure below). The data was then filtered for probe signal intensity: probes which had a p value <0.05 in at least 10% of samples were retained and aggregated to genes using Limma's avereps function. Linear models made using Limma version 3.34.9 were used to assess differential gene expression (6, 7). The empirical Bayes method was employed to moderate the standard errors of the estimated fold-changes (6), and the *arrayWeights* function (8) measured how well the expression values followed the linear model. Correction for multiple testing was done by the Benjamini-Hochberg method; taking an adjusted p value of <0.05 (9).



Gene expression analysis. Quality control and data pre-processing from 2 datasets for gene expression in CD4+T-cell.

Appendix 7

Supplementary method: *Subset phenotyping by flow cytometry*

Flow cytometry was performed using standard cell surface staining protocol using fresh EDTA blood, following red cell lysis. Naïve CD4+T-cells were gated based on the expression of CD3/CD4/CD45RA/CD45RO (as described above). The expression of CD4, IL-6R (CD126 clone M5, BD), IL-2R (CD25 clone 2A3, BD), CXCR4 (CD184 clone 12G5, BD), IL-7R (CD127 clone R34.34, Beckman Coulter)) was measured on naïve CD4+T-cells using Mean Fluorescence Intensity (MFI). Expression of CD62L (clone 145/15, Miltenyi) was either positive or negative and the percentage of CD3+/CD4+/CD45RA+/CD62L- cells was recorded.

Statistical Analysis

Data were displayed as box-plot (ELISA or MFI). Non-parametric Mann-Whitney *U*-test was performed on data comparing HC and RA. Statistical analysis was performed in SPSS V24.

Appendix A8
List of candidate CpGs from 3 strategies for RA classification biomarker

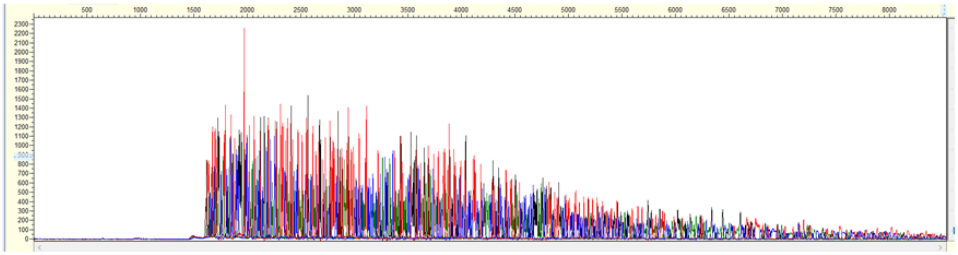
Strategy 1	Probe ID	Nearest gene symbol	Gene name
	cg27183791	ANKRD11	ankyrin repeat domain 11
	cg05299836	BCKDK	branched chain keto acid dehydrogenase kinase
	cg15350899	BCL9L	BCL9 like
	cg01260502	GIMAP7	GTPase, IMAP family member 7
	cg14665366	GPRIN3	GPRIN family member 3
	cg13681468	GPRIN3	GPRIN family member 3
	cg20105257	HLA-E	major histocompatibility complex, class I, E
	cg04618171	HPCAL1	hippocalcin like 1
	cg16379091	IFITM1	interferon induced transmembrane protein 1
	cg03718883	INS-IGF2	INS-IGF2 readthrough
	cg02231590	ITM2C	integral membrane protein 2C
	cg12669088	KRAS	KRAS proto-oncogene, GTPase
	cg05246522	KSR1	kinase suppressor of ras 1
	cg05784862	KSR1	kinase suppressor of ras 1
	cg00759807	LOC100287036	
	cg23660197	MICB	MHC class I polypeptide-related sequence B
	cg20703928	NCK2	NCK adaptor protein 2
	cg17851795	PBX2	PBX homeobox 2
	cg23149454	PDE2A	phosphodiesterase 2A
	cg25256924	PTPRCAP	protein tyrosine phosphatase receptor type C associated protein
	cg03050965	S1PR1	sphingosine-1-phosphate receptor 1
	cg14885762	SEPT9	septin 9
	cg00576086	TERT	telomerase reverse transcriptase
	cg17741993	TNF	tumor necrosis factor
	cg06813419	TRAF5	TNF receptor associated factor 5
	cg01105418	ZBTB18	zinc finger and BTB domain containing 18
Strategy 2	Probe ID	Nearest gene symbol	Gene name
	cg01974478	AP5Z1	adaptor related protein complex 5 subunit zeta 1
	cg12047375	ATP6V1H	ATPase H+ transporting V1 subunit H
	cg13400249	RERE	arginine-glutamic acid dipeptide repeats
	cg04162316	RPTOR	regulatory associated protein of MTOR complex 1
	cg24462702	CD40LG	CD40 ligand
Strategy 3	Probe ID	Nearest gene symbol	Gene name
Result 2	cg02835823	IRF8	interferon regulatory factor 8
Result 3	cg13064571	C8orf44	chromosome 8 open reading frame 44
	cg17002328	CCDC88C	coiled-coil domain containing 88C
	cg05903736	HDAC4	histone deacetylase 4
	cg15058210	HDAC4	histone deacetylase 4
	cg15978561	HDAC4	histone deacetylase 4
	cg02835823	IRF8	interferon regulatory factor 8
	cg12054453	MIR21	microRNA 21
	cg16936953	MIR21	microRNA 21
	cg16853860	PSMB9	proteasome 20S subunit beta 9
	cg20793665	PTMA	prothymosin alpha
	cg22077313	S100P	S100 calcium binding protein P
	cg23458168	ZNF536	zinc finger protein 536
Result 4	cg01106881	DNPEP	aspartyl aminopeptidase
	cg02835823	IRF8	interferon regulatory factor 8
	ch.2.207814544R	KLF7	Kruppel like factor 7
	ch.13.39564907R	LINC00332	long intergenic non-protein coding RNA 332
	cg24174557	MIR21	microRNA 21
	cg26427498	NAMPT	nicotinamide phosphoribosyltransferase
	ch.2.105901354F	NCK2	NCK adaptor protein 2
	cg03206537	NEURL2	neuralized E3 ubiquitin protein ligase 2
	cg08752433	PPTC7	protein phosphatase targeting COQ7
	cg20388732	STAT5A	signal transducer and activator of transcription 5A
	cg00004B5:K2566	ZBTB17	zinc finger and BTB domain containing 17

Appendix 9

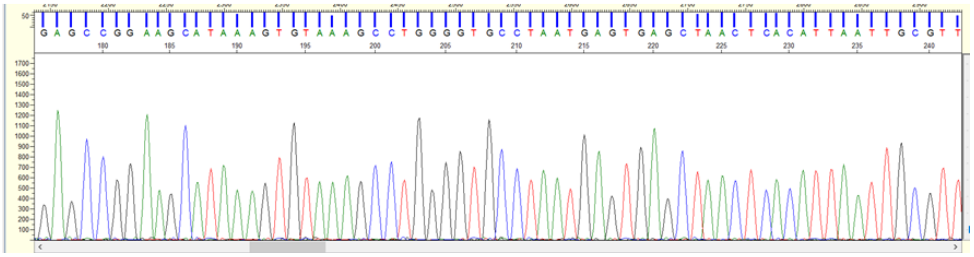
Bisulfite sequencing

A good sequencing example form the control sequencing reaction using the company template DNA and primer.

Raw intensity



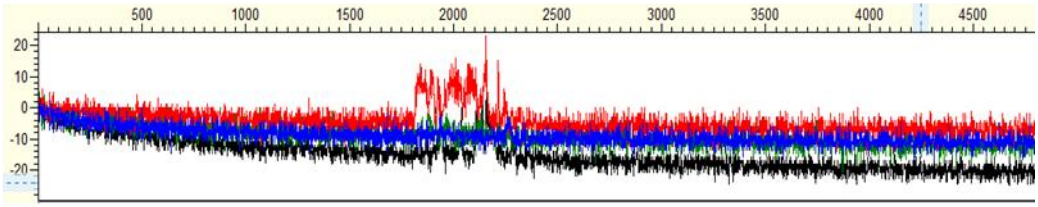
Electropherogram



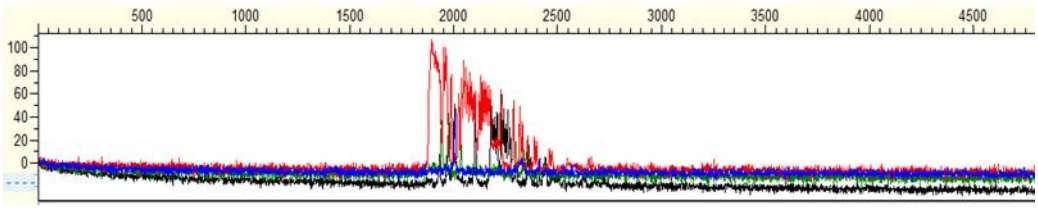
IFITM1 Sequencing reaction optimisation: primer concentration

Sequencing raw intensity results using different primer concentration. Low signal intensity is observed and the signal present only in the beginning part of the sequence. Reducing primer concentration slightly increased the sequencing signal intensity.

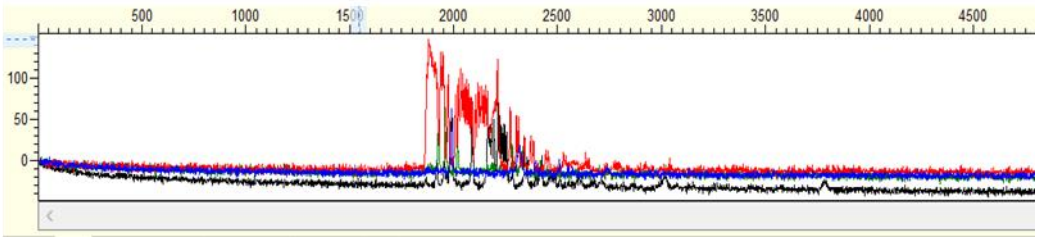
Primer at 160 nM



Primer at 80 nM



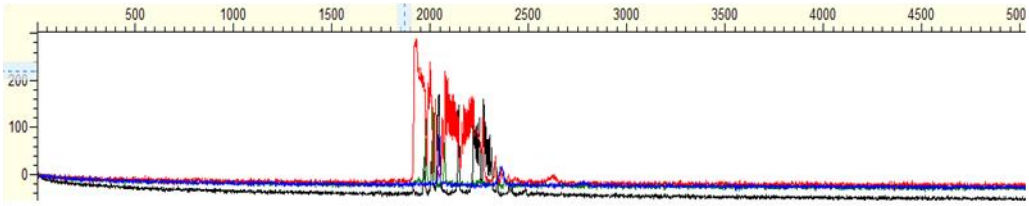
Primer at 40 nM



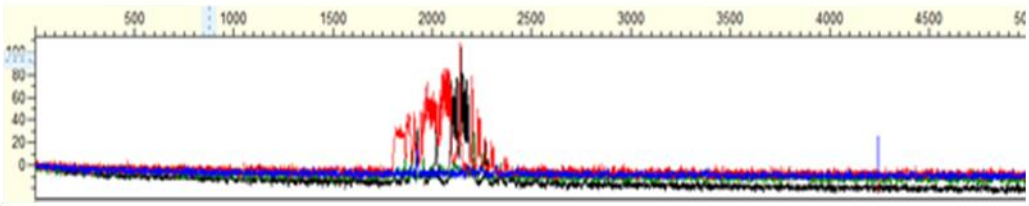
IFITM1 Sequencing reaction optimisation: amount of template

Sequencing raw intensity results using different amount of template. Low signal intensity is observed and the signal present only in the beginning part of the sequence. Reducing template input slightly increased the sequencing signal intensity.

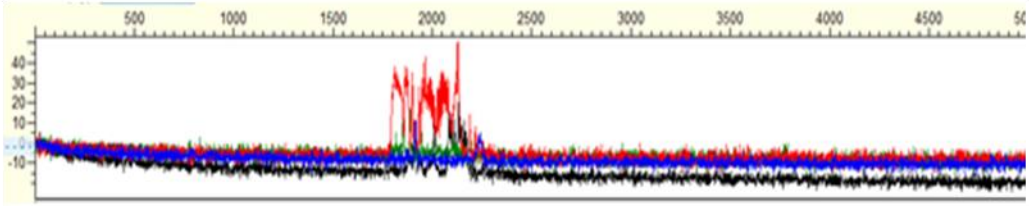
DNA template 0.5 ul



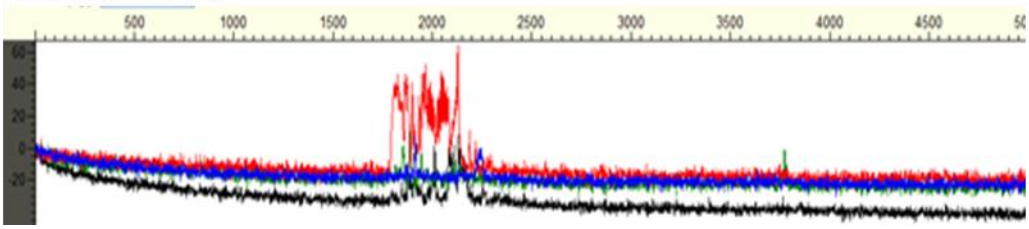
DNA template 1 ul



DNA template 2 ul



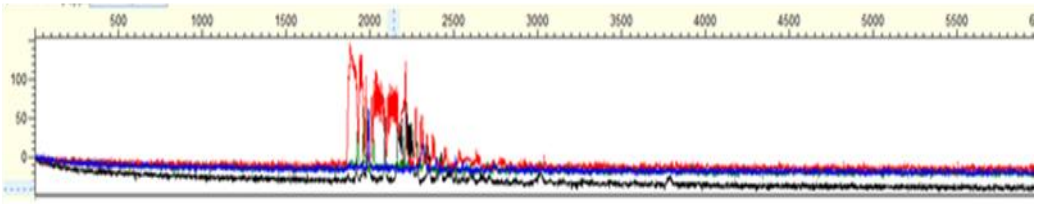
DNA template 4 ul



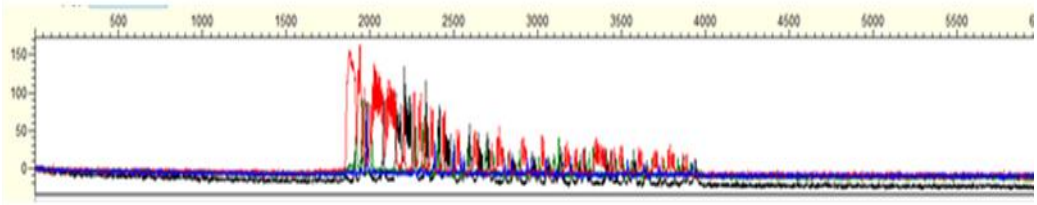
IFITM1 Sequencing reaction optimisation: enzyme concentration

Sequencing raw intensity results using different enzyme concertation. Increasing reaction enzyme notable help increase the signal intensity. The whole length of the PCR product could be sequenced at higher enzyme concentration.

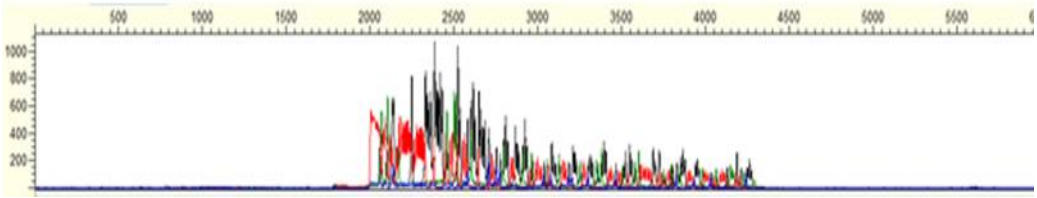
Reaction enzyme at 0.0625 X



Reaction enzyme at 0.1250 X



Reaction enzyme at 0.25 X

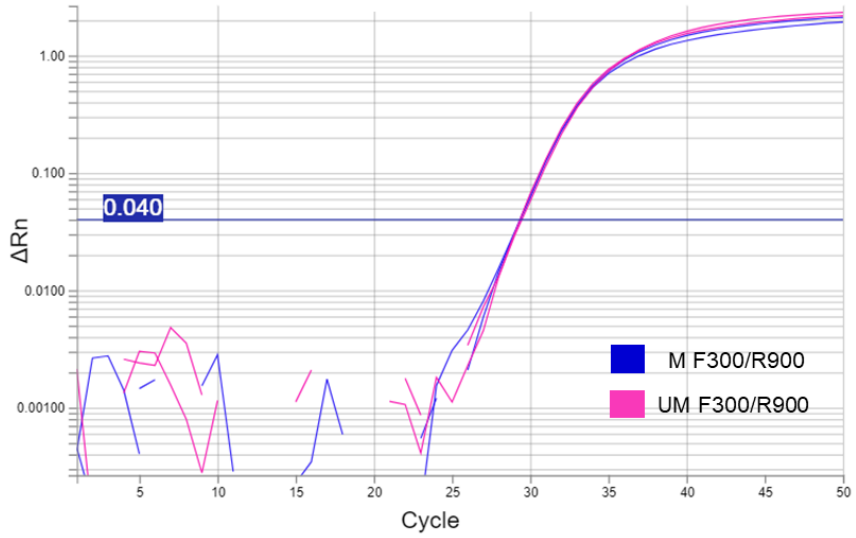


Appendix 10

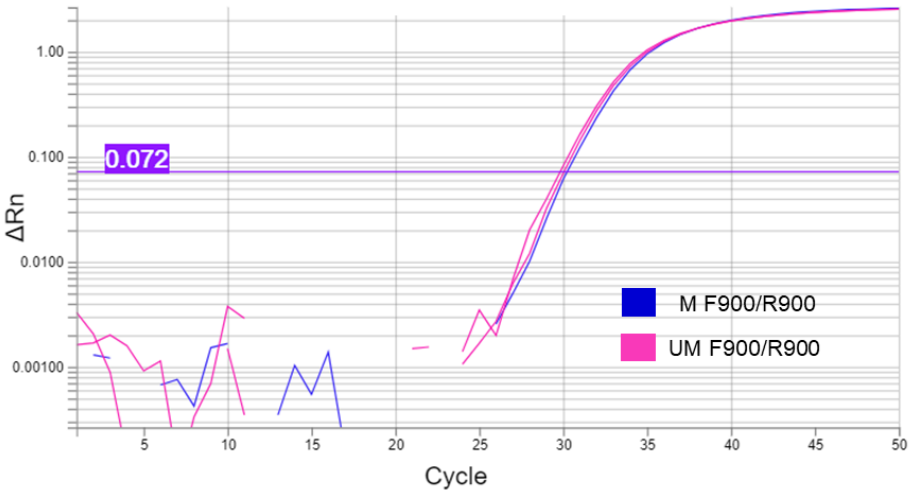
TaqMan qMSP assay development

Amplification plot of GAPDH, ACTB TaqMan qMSP assay using of 100% methylated and un-methylated DNA. The methylation independence for the internal control assay; GAPDH and ACTB assay was confirmed.

GAPDH



ACTB

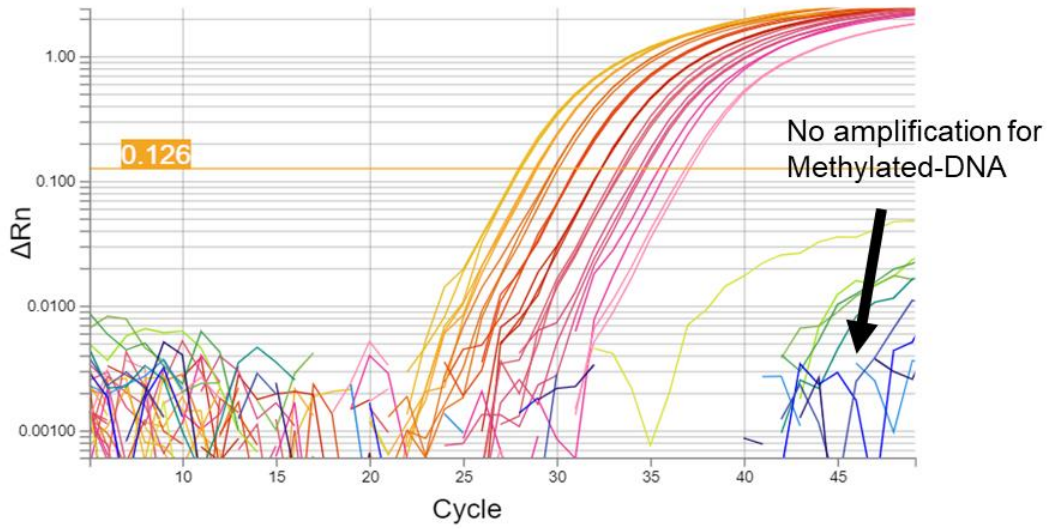


Appendix 11

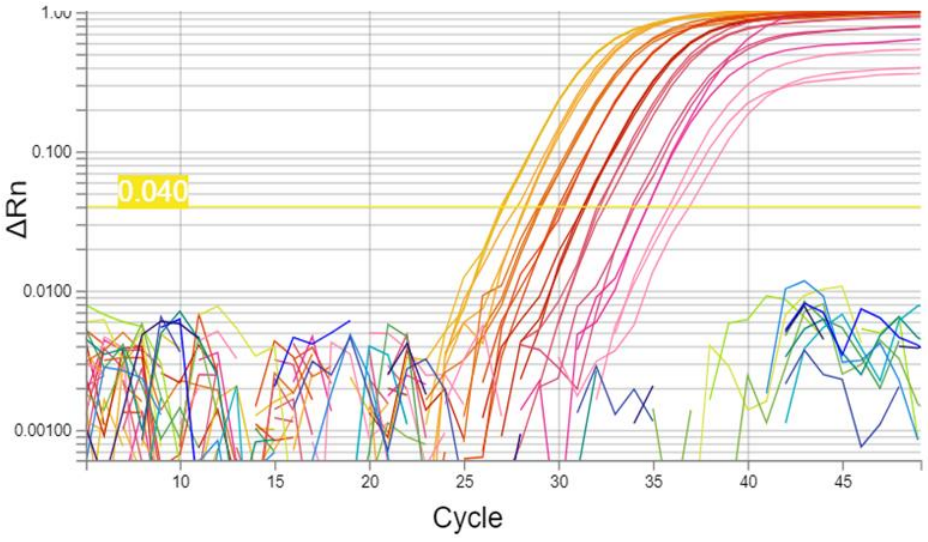
TaqMan qMSP assay development

Amplification plot of HDAC4, TNF, GAPDH, ACTB TaqMan qMSP assay using serial dilution of 100% methylated and un-methylated DNA. The colours from yellow to pink corresponding to 0.2 ng to 50 ng of methylated DNA. and the colour form green to blue corresponding to the unamplifying un-methylated DNA.

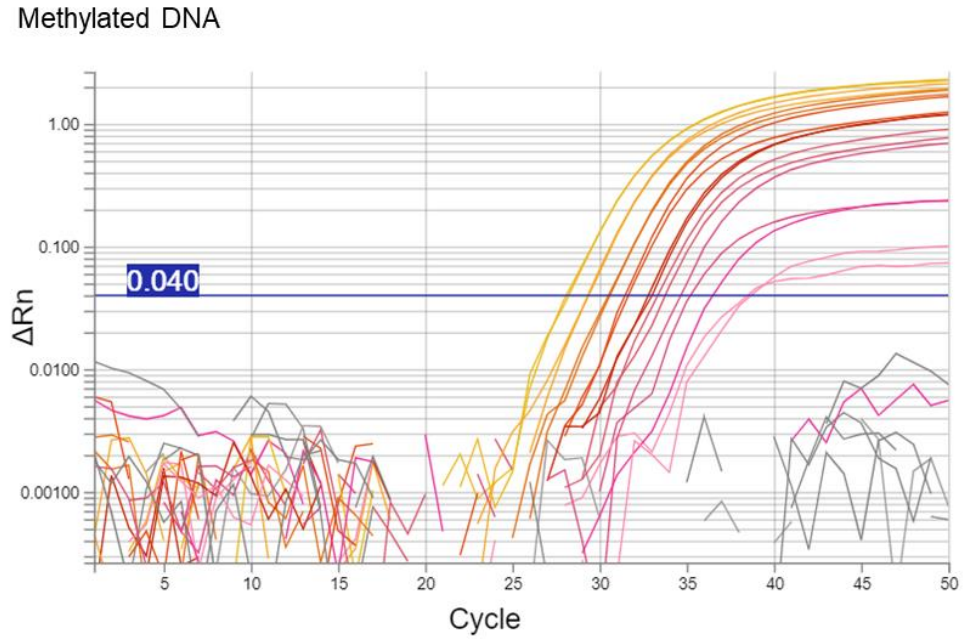
TNF

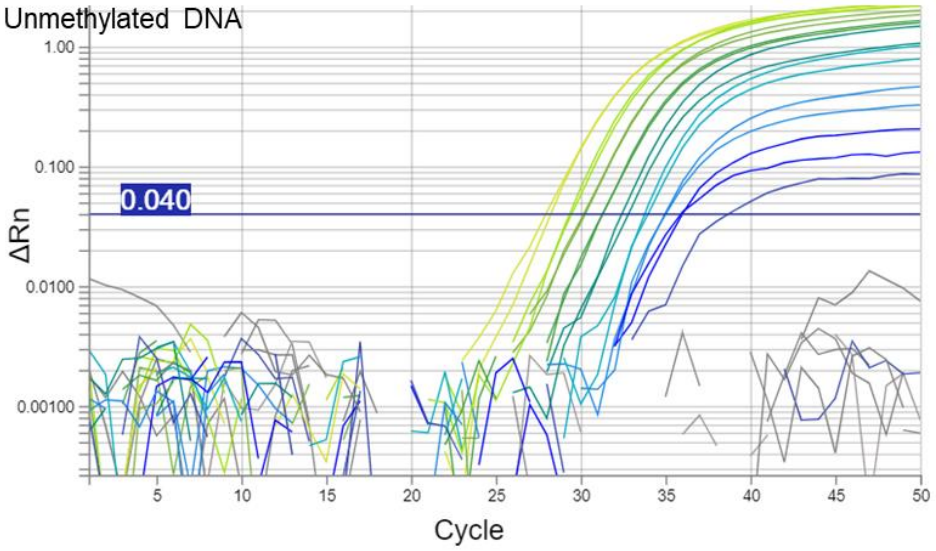


HDAC4



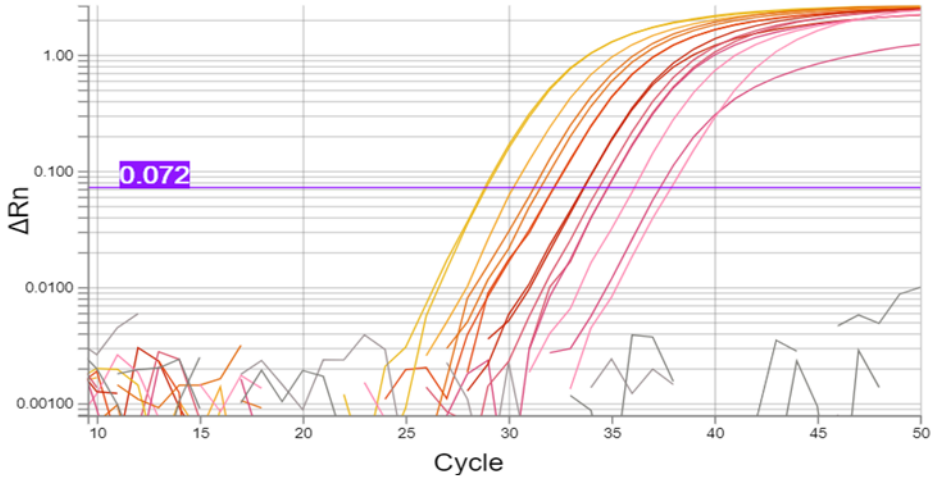
GAPDH


Methylated DNA
1.00
0.100
0.040
ΔRn
0.0100
0.00100
5
10
15
20
25
30
35
40
45
50
Cycle

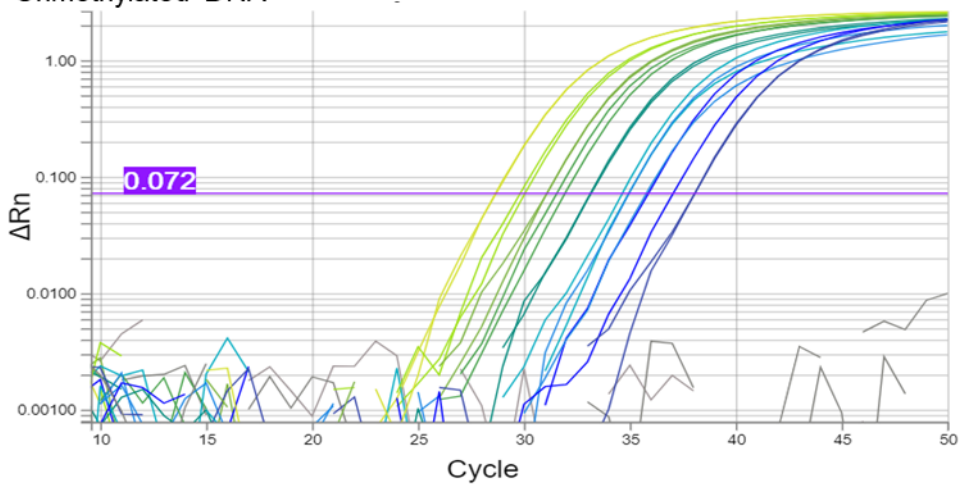

Unmethylated DNA
1.00
0.100
0.040
ΔRn
0.0100
0.00100
5
10
15
20
25
30
35
40
45
50
Cycle

ACTB

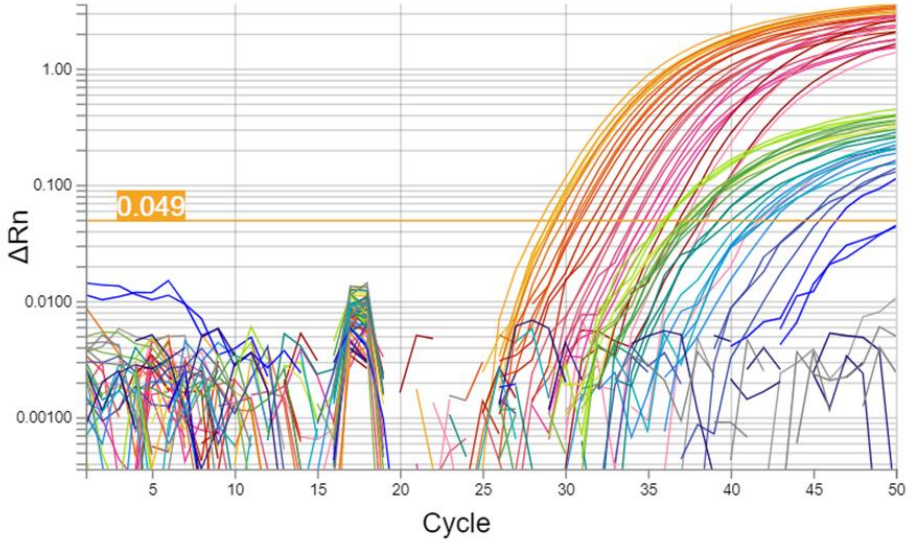
Methylated DNA



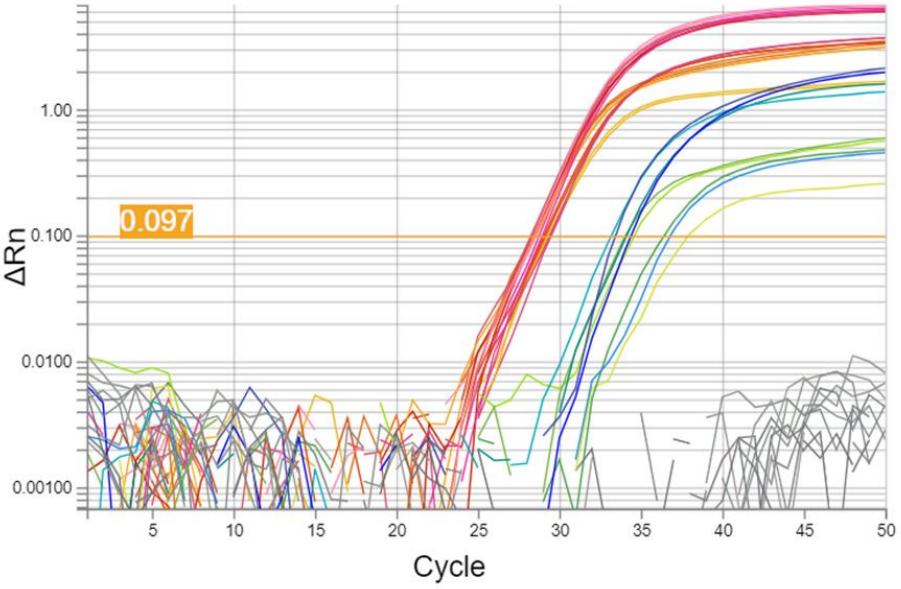
Unmethylated DNA



IRF8


1.00
0.100
0.0100
0.00100
ΔRn
0.049
5
10
15
20
25
30
35
40
45
50
Cycle

IRF8 V.2


1.00
0.100
0.0100
0.00100
ΔRn
0.097
5
10
15
20
25
30
35
40
45
50
Cycle

Appendix references

1. Neogi T, Aletaha D, Silman AJ, Naden RL, Felson DT, Aggarwal R, et al. The 2010 American College of Rheumatology/European League Against Rheumatism classification criteria for rheumatoid arthritis: Phase 2 methodological report. Arthritis Rheum. 2010;62(9):2582-91.
2. Pratt AG, Swan DC, Richardson S, Wilson G, Hilkens CM, Young DA, et al. A CD4 T cell gene signature for early rheumatoid arthritis implicates interleukin 6-mediated STAT3 signalling, particularly in anti-citrullinated peptide antibody-negative disease. Annals of the rheumatic diseases. 2012.
3. Gruden K, Hren M, Herman A, Blejec A, Albrecht T, Selbig J, et al. A "crossomics" study analysing variability of different components in peripheral blood of healthy caucasoid individuals. Plos One. 2012;7(1):e28761.
4. Lin SM, Du P, Huber W, Kibbe WA. Model-based variance-stabilizing transformation for Illumina microarray data. Nucleic Acids Res. 2008;36(2):e11.
5. Du P, Kibbe WA, Lin SM. lumi: a pipeline for processing Illumina microarray. Bioinformatics. 2008;24(13):1547-8.
6. Ritchie ME, Phipson B, Wu D, Hu YF, Law CW, Shi W, et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Research. 2015;43(7).
7. Phipson B, Lee S, Majewski IJ, Alexander WS, Smyth GK. Robust Hyperparameter Estimation Protects against Hypervariable Genes and Improves Power to Detect Differential Expression. Ann Appl Stat. 2016;10(2):946-63.
8. Ritchie ME, Diyagama D, Neilson J, van Laar R, Dobrovic A, Holloway A, et al. Empirical array quality weights in the analysis of microarray data. Bmc Bioinformatics. 2006;7.
9. Benjamini Y, Hochberg Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. Journal of the Royal Statistical Society Series B (Methodological). 1995;57(1):289-300.