Identifying the molecular signatures that shape the course of synovial pathology in inflammatory arthritis.

CARDIFF
UNIVERSITY

PRIFYSGOL
CAERDYDD

A Thesis submitted to Cardiff University in accordance to the requirements for the degree of Doctor of Philosophy in the School of Medicine

By

Benjamin C. Cossins BSc (Hons) MSc

November 2020

Summary.

Advances in precision medicine offer exciting opportunities to improve healthcare provision and clinical decision-making. Here, developments in diagnostic capabilities provide greater insights into the mechanisms of disease progression and allow the stratification of patients for the selection of therapies for optimal treatment. Innovations in precision medicine, therefore, contribute to improved clinical outcomes, a patient's quality of life, and health economics. Experiment presented in this investigated the development of bioinformatic tools that could be used to stratify patients based on transcriptomic data derived inflamed tissues. To support this approach, I used open access repository datasets from patients with rheumatoid arthritis.

Rheumatoid arthritis (RA) is a chronic and systemic autoimmune disease that affects around 1% of the adult population. Here, inflammation of the joint (synovitis) drives disease progression and irreversible joint damage. The clinical presentation of synovitis is highly heterogeneous with distinct histological features that affects the response commonly used therapeutics (e.g., biological drugs against cytokines).

Examination of synovial histopathology reveals three forms of the disease termed Follicular – with extensive infiltration and the presence of lymphoid aggregates; Diffuse – extensive infiltration but with relatively few B cells; and Pauci-immune – which is driven by the stromal tissue compartment. Using transcriptomic data from each of these pathologies I designed and validated a disease classifier that supports the stratification of disease and the interrogation of results from independent patient cohorts where batch effects often restrict interpretations. Thus, I now present a tool that allows discrimination of these pathologies according to synovial transcriptomic data.

Results presented in this thesis identified two gene signatures that perform well as identifiers of follicular and pauci-immune synovitis. The characterisation of diffuse synovitis is, however, more challenging and the application of the disease classifier tools showed that this form of pathology comprises a spectrum of sub-pathologies that require further characterisation. In an extension of these studies I further show how these bioinformatic tools may be used to record patient responses to biological drug therapy and unearth the biological signal pathways responsible for disease progression. Whilst these studies have focussed on RA as a case study, the methodologies are disease agnostic, and offer exciting opportunities for additional applications in other disease settings.

Acknowledgements.

The past four and a half years have been amazing, whilst there are have been highs and lows, I have thoroughly enjoyed my time in the lab.

First, I would like to thank my supervisor, Professor Simon Jones, for providing me with the opportunity to do this PhD, his infinite patience and support that has encouraged and dragged me through to complete my writing. I would like to thank my co-supervisors: Dr Gareth Jones for all his time helping me with the mouse models and discussions for my numerous questions and problems throughout this project, and Professor Nigel Williams for his encouragement and insights on my work. I would also like to thank Professor Valerie O'Donnell for introducing and recommending me for the opportunity to do this PhD.

I would like to thank the rest of the lab for being amazingly supportive and welcoming, and for the many Friday lunches we shared. I would like to thank Dr Anna Cardus for working extensively with me in from the very first day, introducing me to ChIP-seq, and for looking after my geckos more than once. I would like to thank Dr Jason Twohig for essentially adopting me into his family, and for countless hours debating analysis methods. Further thanks extended to Dr Xiao "Tommy" Liu for assisting me early on with working with mice, to Dr Rob Jenkins for all of the laughs, and not killing me for moving the occasional pencil on his desk, and to Dr James Burston for all the discussions that helped me focus my writing.

I would also like to thank my fellow PhD students for all their assistance in the lab. Specifically, Dr David Hill for challenging me to up my cake game, Dr Alicia Derrac-Soria for helping me with all the English questions, Dr Katie Sime for making lunches an interesting time, and Aisling Morrin for all the laughs. I bring special mention to Dr Javier Uceda Fernández for being such an enthusiastic person and is sorely missed.

I would also like to thank Dr Robert Andrews for all the hours of support that have made me a better bioinformatician. I would like to thank Dr Amanda Tonks and her PGR team for all the support they have provided me over this PhD.

Finally, I would like to thank my family for their enthusiasm and support throughout; especially my sister, Lani (BSc, MSc), for all the time spent proofreading some of the worst first drafts before Simon even had to suffer them.

Publications and presentations.

Publications.

Twohig, J.P., Cardus Figueras, A., Andrews, A., Wiede, F., **Cossins, B.C.,** Derrac Soria, A., Lewis, M.J., Townsend, M.J., Millrine, D., Li, J., Hill, D.G., Uceda Fernandez, J., Liu, X., Szomolay, B., Pepper, C.J., Taylor, P.R., Pitzalis, C., Tiganis, T., Williams, N.M., Jones, G.W., & Jones, S.A. **Activation of naïve CD4+ T cells re-tunes STAT1 signalling to deliver unique cytokine responses in memory CD4+ T cells.** Nat Immunol 20, 458–470 (2019)

Khalid, U., Jenkins, R.H., Andrews, R., Pino-Chavez, G., **Cossins, B.C.,** Chavez, R., Bowen, T., & Fraser, D.J. **Determination of a microRNA signature of protective kidney ischemic preconditioning originating from proximal tubules.** Manuscript submitted for publication.

Presentations.

Cossins, B.C, Andrews, R., Twohig, J., Williams, N., Jones, G., Jones, S. **Transcriptional profiles of synovial biopsies predict therapeutic response to biologics in rheumatoid arthritis.** Poster presented at: 6th Annual meeting of the International Cytokine & Interferon Society, 2018, Boston. Abstract P002.

Abbreviations.

ACPA	Anti-Citrullinated Proteins Antibodies
ACR	American College of Rheumatology
AIA	Antigen-induced arthritis
ATAC-seq	Assay for Transposase-Accessible Chromatin-sequencing
AUC	Area Under the Curve
BCL3	B-cell lymphoma 3-encoded protein
BCL6	B-cell lymphoma 6 protein
BEScore	Batch Effect Score
BWA	Burrows-Wheeler Aligner
CARD1	Caspase Recruitment Domain Family Member 11
CD	Cluster of Differentiation
CD4	Cluster of Differentiation 4
CDAI	Clinical Disease Activity Index
CFA	Complete Freud’s Adjuvant
ChIP	Chromatin Immunoprecipitation
ChIP-seq	Chromatin ImmunoPrecipitation-sequencing
CLiP-seq	Cross-Linking ImmunoPrecipitation-sequencing
CRP	C-Reactive Protein
CSV	character separated value
CTLA4	Cytotoxic T-Lymphocyte Associated Protein 4
DAS28	Disease Activity Score 28
DMARD	Disease-Modifying Anti-Rheumatic Drugs
DNA	Deoxyribonucleic acid
EBI	European Bioinformatics Institute
EBV	Epstein-Barr Virus
EDTA	Ethylene glycol-bis(2-aminoethylether)-N,N,N′,N′-tetraacetic acid
EGTA	Ethylenediaminetetraacetic acid
HEPES	2-[4-(2-hydroxyethyl)piperazin-1-yl]ethanesulfonic acid
IL4R	Interleukin-4 receptor
ESR	Erythrocyte Sedimentation Rate
EULAR	European League Against Rheumatism
FAIR	Findability, Accessibility, Interoperability, and Reusability
FDR	False Discovery Rate
FPKM	Fragments Per Kilobase of transcript per Million mapped reads
GAS	Gamma interferon activation site
GEO	Gene Expression Omnibus
HAQ	Health Assessment Questionnaire disability index
HER2	Human Epidermal Growth Factor Receptor 2
HLA-DRB1	HLA class II histocompatibility antigen, DRB1 beta chain
I.A.	Intra-Articular
I.P.	IntraPeritoneally
IAM	Intercellular Adhesion Molecule 1
ICOS	Inducible T Cell Costimulator

IFN	Interferon
IgG	Immunoglobulin G
IL	Interleukin
IPA	Ingenuity Pathway Analysis
IPDB	IP Dilution Buffer
IRF1	Interferon regulatory factor 1
LAT	Linker For Activation Of T Cells
LILRA3	Leukocyte Immunoglobulin Like Receptor A3
MACS	Model-based Analysis of ChIP-seq
MAIME	Minimum Information About a Microarray Experiment
mBSA	methylated Bovine Serum Albumin
NCBI	National Center for Biotechnology Information
ND	No Disease
NICE	National Institute for Health and Care Excellence
NK cell	Natural Killer cell
NLB	Nuclear Lysis Buffer
NP40	Nonyl Phenoxypolyethoxylethanol
OA	OsteoArthritis
PAM	Prediction Analysis for Microarrays
PASII	Patient Activity Scale-II
PAVIS	Peak Annotation and VISualisation
PBS	Phosphate Buffered Saline
PCA	Principle Component Analysis
PCR	Polymerase Chain Reaction
PEAC	Pathobiology of Early Arthtitis Cohort
PMSF	PhenylMethaneSulfonyl Fluoride
PRKCB	Protein Kinase C Beta
PTPN22	Protein tyrosine phosphatase, non-receptor type 22
PTX	Pertussis Toxin
qPCR	Quantitative Polymerase Chain Reaction
R4RA	Response - Resistance to Rituximab versus Tocilizumab in RA
RA	Rheumatoid Arthritis
RAPID3	Routine Assessment of Patient index 3
RF	Rheumatoid Factors
RMA	Robust Multi-Array
RNA	Ribonucleic acid
RNA-Seq	RNA-sequencing
ROC	Reciever Operator Characteristics
S.C.	SubCutaneous
SDAI	Simplified Disease Activity Index
SDS	Sodium Dodecyl Sulfate
SNP	Single Nucleotide Polymorphisms
sPLS-DA	sparse Partial Lease Squares Discriminatory Analysis
STAT	Signal Transducer and Activator of Transcription
STRAP	Stratification of Biologic Therapies for RA by Pathobiology

TCR	T Cell Receptor
TNF	Tumour Necrosis Factor
TPM	Transcripts Per Million
TSS	Transcription Start Site
VAS	Visual Analogue Scale
WT	Wild Type
ZAP70	Zeta Chain Of T Cell Receptor Associated Protein Kinase 70
ZEB1	Zinc finger E-box-binding homeobox 1

Table of Contents

List of Figures.

List of Tables.

1. General introduction.

1.1. Background.

Initiated in 1990, the Human Genome Project was an international investment designed to determine the DNA sequence of the entire human genome. Since the publication of the first datasets in 2003 the subsequent two-decade has seen an explosion in our understanding of human genetics in health and disease and has pathed the way for significant advances in both computational methodologies and the development of sequencing technologies(1–3). Following the mapping of the human genome, various technologies have been developed that allow a genome-wide interrogation of genetic determinants of disease susceptibilities and investigations into the epigenetic mechanisms affecting gene regulation. Consequently, researchers interested in genetics and functional genomics now benefit from a range of technologies that couple traditional biochemical methods with whole-genome next-generation sequencing capabilities. These include the identification of Single Nucleotide Polymorphisms (SNPs) that predict genetic susceptibility for a disease or a biological trait within a population and studies of gene regulation. Here, the application of RNA-sequencing (RNA-seq), chromatin immunoprecipitation-sequencing (ChIP-seq), cross-linking immunoprecipitation-sequencing (CLiP-seq), assay for transposase-accessible chromatin-sequencing (ATAC-seq) and others have enhanced our ability to identify genes and pathways associated with aberrant behaviour. In parallel with these technology advances, we have also seen a significant expansion in computational capabilities and biologists with expertise in bioinformatics, biostatistics and mathematical modelling are increasingly used to support fundamental discovery science and clinical studies.

1.1.1. Big data.

Advances in sequencing technology have reduced the time required to sequence the human genome. While the original human genome project took 13 years to complete, the human genome can now be sequenced within 24 hours using the latest technologies (4). Genetic sequencing through The Sanger Institute produced approximately 1 petabyte of genomic data up to 2012. In 2019, this same amount of data was generated every 35 days (5). During this time, the overall cost of sequencing has dropped considerably and the ability to multiplex samples for analysis has made the technology increasingly accessible. For example, studies performed by my laboratory have almost entirely switched away from the use of quantitative-PCR methods to more holistic RNA-seq, which offers a greater amount of information for a similar cost.

This increase in the amount of available genetic and genomic data is similarly reflected by an increase in the annual number of publications. Here, the number of papers with the mention of “gene expression” in 2019 is double the number of publications per year since the publication of the human genome project (Figure 1.1). Moreover, the number of publications continues to rise year over year. These large datasets, therefore, provide the opportunity to utilise discriminatory analyses and machine learning to identify patterns of behaviour in the data.

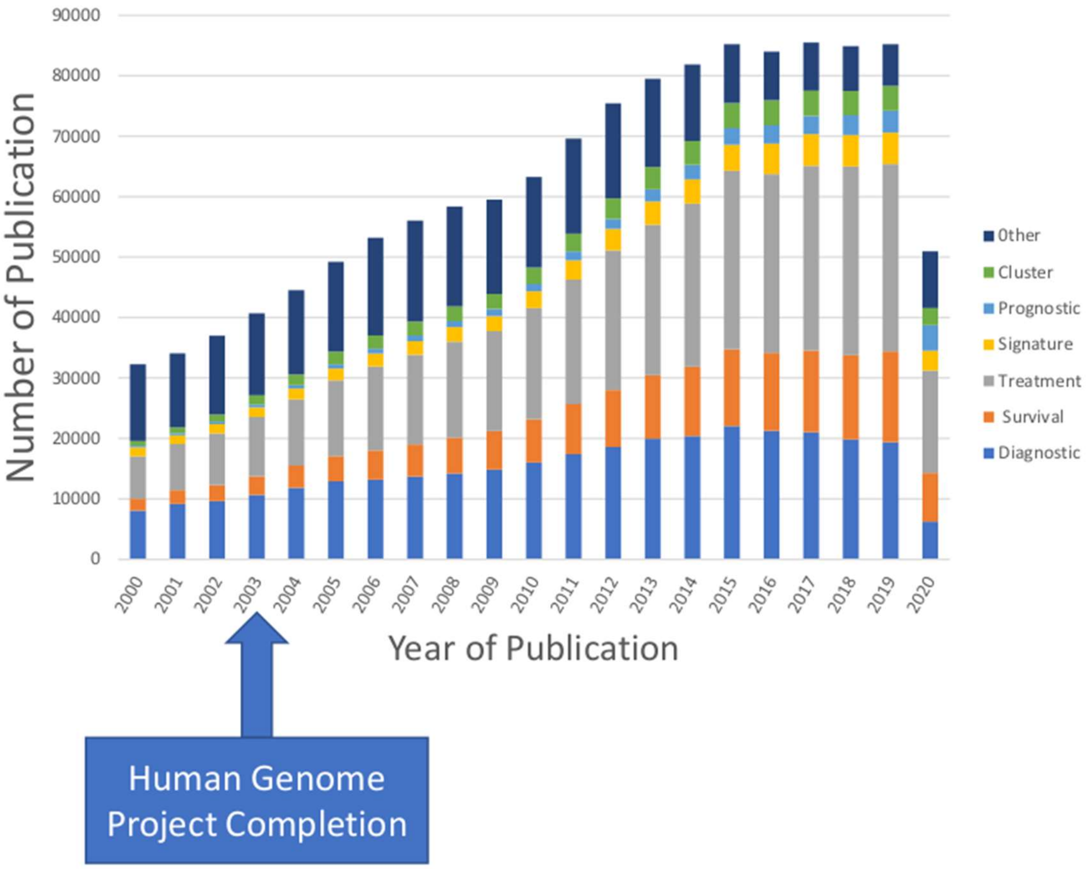


Figure 1.1: Number of publications per year with the term “gene expression”.

This was also expanded to identify papers that included other terms, Signature, Treatment, Survival, Diagnostic, Prognostic.

1.1.2. Computational analyses.

Analysing transcriptomic data can be performed utilising a large range of statistical methodologies. However, these can be broadly classified into three main categories: differential gene expression, clustering, and prediction.

- Clustering analyses do not include *a priori* information. Instead, these methods attempt to discern fundamental similarities between samples that identify common features within the datasets. These often equate to genes or samples that display a common pattern of expression or contribute to a common underlying biological mechanism.

- Differential gene analysis aims to identify genes that are fundamentally different between groups. This analysis necessitating prior information that identifies the groups (although this may be derived from clustering). These types of approaches are often applied to understand the biological significance of gene deletions or responses to interventions that target a specific protein or signalling pathway.
- Predictive analyses can be implemented either diagnostically or prognostically. These analyses use prior knowledge to identify patterns in the data that allow for the discrimination of the groups. Future samples can then be classified based on these predictive features.

These methodologies have demonstrated extensive and increasing usage with gene expression data, as visualised in Figure1.1. These techniques are explored in a little more detail in the next section

1.1.3. Clustering.

Methods designed to allowing clustering are broadly used in various fields of investigation. Clustering is an unsupervised learning methodology(6). However, care must be taken to identify the correct clustering algorithm, as no one algorithm performs well for all problems(7). This thesis makes extensive use of hierarchical clustering and Principle Component Analysis (PCA) to reduce the high dimensionality of the data to a 2D representation of the data to identify groups.

Hierarchical clustering can be approached using two principle methods (8). (A) Agglomerative nesting – where each sample is its own cluster and then iterates to make the two most similar samples is a cluster and progresses to include all samples. (B) Divisive analysis – this method takes the opposite approach and starts with the whole population and splits the data into progressively smaller parts until each sample is distinct. Both of these approaches using statistical assessments to sub-group the original data but have numerous pro's and con's that influence the choice of analysis used. In this thesis the primary algorithms utilised was agglomerative hierarchical clustering using Euclidean distances and Ward linkage method.

PCA is primarily a tool to reduce the dimensionality of the data, reducing thousands of genes to only the few that have the most impact on variance. In this thesis, whilst not applied as a clustering algorithm, PCA allows interrogation of the datasets to determine if variance observed in the samples separates them into meaningful groups.

1.1.4. Differential gene expression.

Differential gene expression is utilised to identify the changes in transcription activity associated with treatment or the phenotype of interest. At a basic level, this can be achieved with simple T-tests, which was often performed when using small numbers of genes. For example, studies of gene regulation performed using polymerase chain reaction (PCR) amplification. On the larger scale of whole exomes, Bayesian approaches have been utilised with microarray and RNA-sequencing. This thesis makes extensive use of limma, which is a package in the R programming language that utilises a parametric empirical Bayesian approach that allows meaningful differentially expressed genes to be observed from relatively few samples(9).

1.1.5. Predictive Modelling.

Predictive models aim to identify the features that discriminate samples based on the outcome and is a similar analysis to that described for differential gene expression. This thesis made use of three algorithms to identify gene signature that discriminates the different pathologies of interest: Shrunken centroids, sparse Partial Lease Squares Discriminatory Analysis (sPLS-DA), and Random Forrest.

1.1.5.1. Shrunken centroids.

The use of shrunken centroids was developed as an approach to stratify different cancer types using microarray data(10). Given the large number of probes on a microarray, it was important to reduce the number of genes to those that contribute the most towards classification. Moreover, this technique was optimised for discriminating multiple classes.

In this approach, the average expression of each gene is calculated for the entire dataset as well as for each of the classes. Subtracting the global centroid from the class-specific centroids allows identification of those genes that are most distinct. Whilst this allows comparison of sample profiles against these classes, it utilises the entire transcriptome to perform stratification. Therefore, the shrunken component of the analysis allows the reduction of the number of genes needed to discriminate the classes. Shrinking is performed by standardisation by the within-class standard deviation, weighting genes by those that have most stable (less variance) expression. A soft thresholding approach allows control of the level of shrinkage, and therefore the number of genes selected for discrimination. 10-fold cross-validation it then used to determine the overall error at each thresholding level. Figure 1.2 illustrates the centroids of the data, effects of shrinkage, shrinkage threshold, and the resulting genes selected that discriminates the classes.

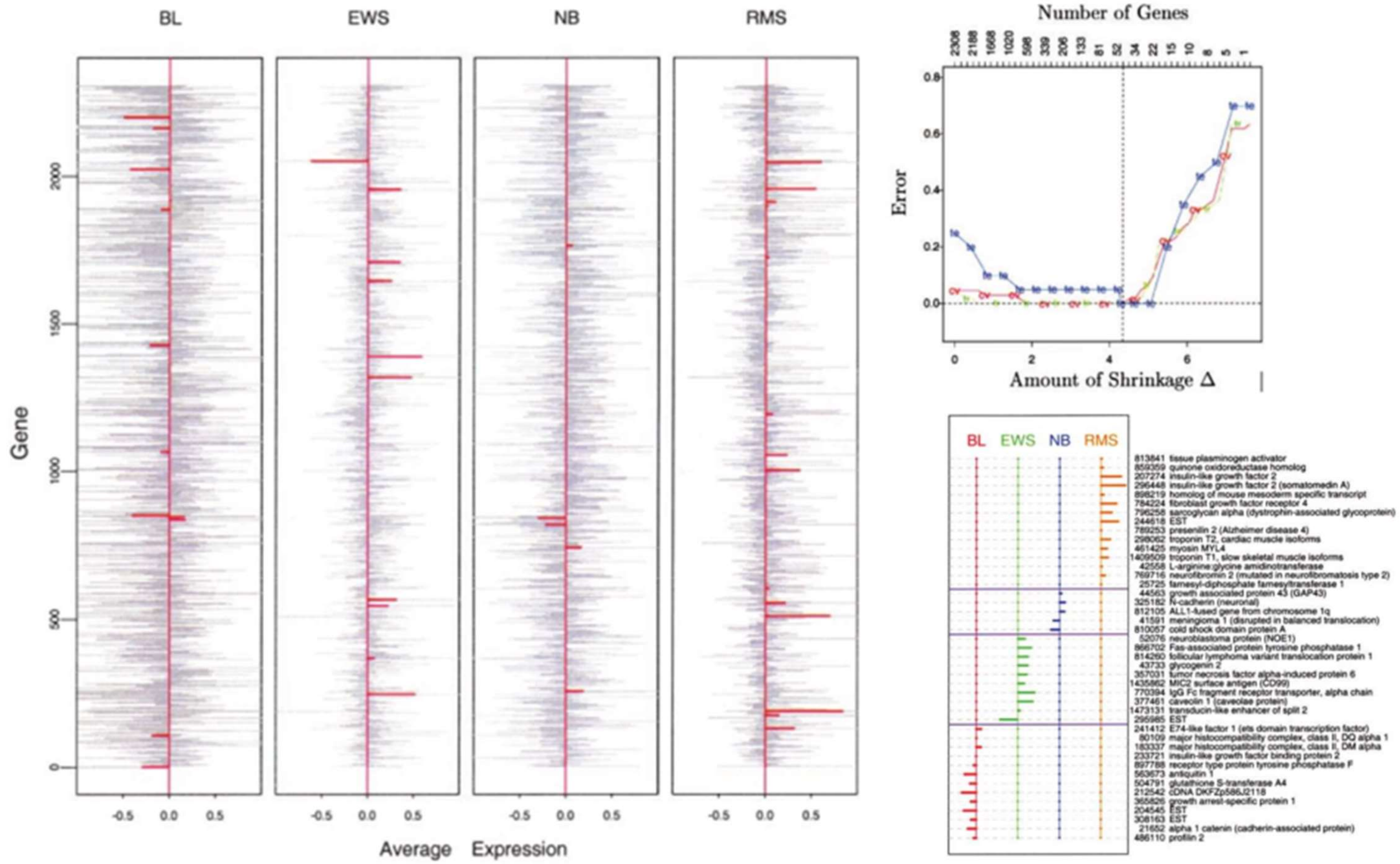


Figure 1.2: Example of shrunken centroids classification of 4 classes of cancer.

***A** Grey bars show shows the centroid values for each class, red shows the denoised profile resulting from the application of shrinkage. **B** Highlights the number of genes and general error rate at each threshold. **C** the resulting 43 genes that differentiate the 4 cancer classes, and the contribution to each centroid. Figure adapted from Tibshirani et al. (2002). Diagnosis of multiple cancer types by shrunken centroids of gene expression. PNAS*

1.1.5.2. Random Forest.

Its name comes from the creation of numerous decision trees, resulting in a "forest", from which a majority vote allows the identification of the most important features. Figure 1.3 illustrates a very simple decision tree, that allows the stratification of three groups using two rules. In the random forest approach, multiple decision trees are generated by randomisation of variables and rules, and the importance of these is calculated by majority vote.

In the context of disease stratification, we need to learn the pattern of expression associated with the disease classes. Using multiple iterations, it is then possible to rank genes in terms of their contribution, and therefore create a classifier model based of those genes most important.

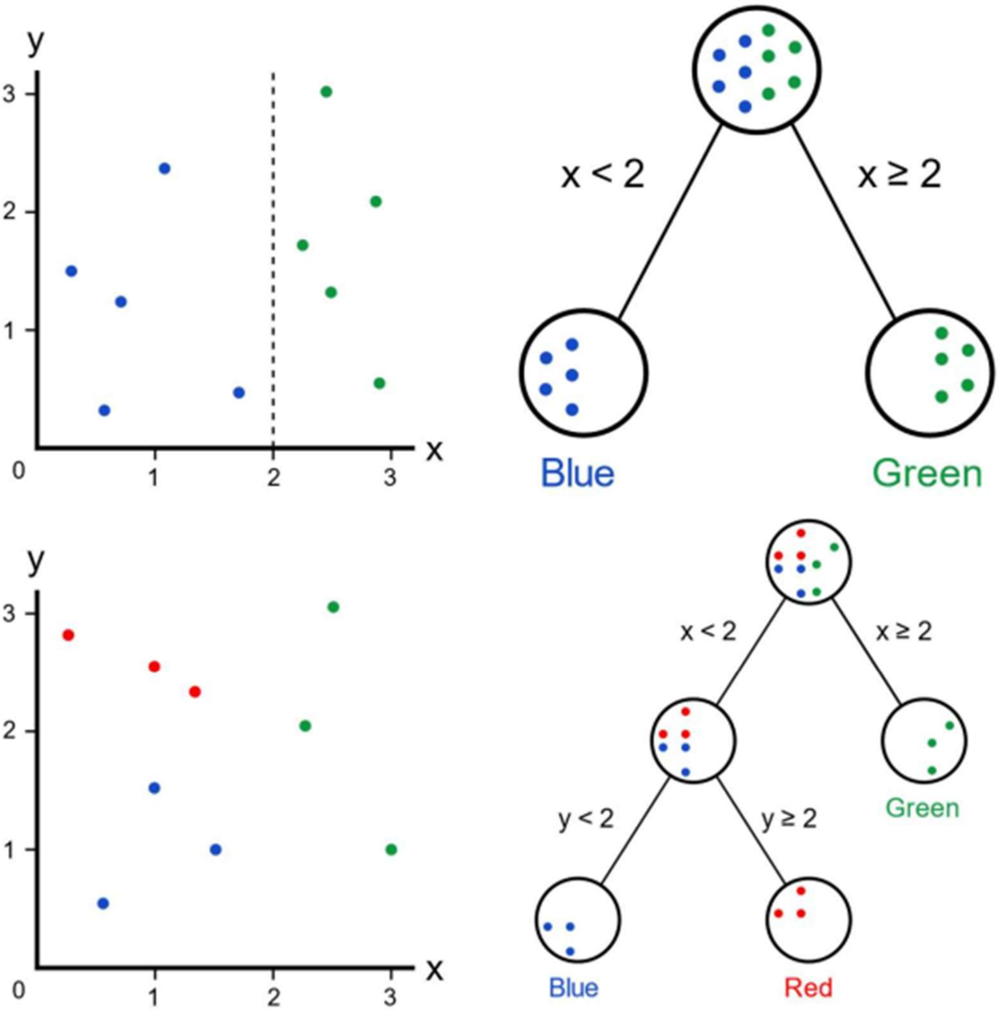


Figure 1.3: Example of simple decision trees.

*(**A**) A simple dataset with 2 classes, creating a simple rule (**B**) of x<2 = blue, x>2 green, however, this no longer works with a slightly more complicated dataset with 3 classes (**C**) in this case we need to incorporate additional rules (**D**).*

1.1.5.3. Partial Least-Squares Discriminatory Analysis.

Partial least-squares discriminatory analysis is a multivariate dimensionality-reduction tool, it can be thought of as a 'supervised' PCA. The difference is that it reduces the dimensions, but with the full awareness of the class labels(11).

The technique can be utilised for feature selection and classification(12), although it is prone to abuse though overfitting. In principle component analysis eigenvectors are created based on variance, and groups are identified by how they cluster after dimensional reduction. In partial least squares, we select for the eigenvector that allows the segregation of the groups(13).

Given that only a small number of features (genes) may be responsible for driving a biological event, we can adapt the technique to incorporate a sparsity assumption. Doing so we can assign the number of features to each component and assess the ability of the model to stratify the groups. This is where the challenge appears, what are the optimal number of features to incorporate in each component. To address this, a bootstrapping method performs multiple rounds of sPLS-DA using different features to identify the core set of features that are consistent amongst the models(14)

1.2. Precision medicine.

Advances in clinical medicine and clinical innovation have seen a significant move towards a more personalised approach to patient treatment. Here, precision medicine methods are increasingly used to understand the pathways driving disease and to identify optimal treatments that have the best efficacy for a patient or patient group. To support these decision-making processes, biological, clinical and imaging biomarkers are often combined to improve the diagnostic and prognostic interpretation and the stratification of patients for optimal therapy. A famous example of this would be the application of the biological drug trastuzumab, which is a monoclonal antibody used the treatment of HER2-positive breast cancer(15).

As a model disease, Rheumatoid Arthritis provides an ideal target for precision medicine. It is a common autoimmune disease that affects around 1% of the adult population, and patient response to therapy is heterogeneous. Random controlled trials have proven the efficacy of conventional disease-modifying anti-rheumatic drugs (DMARD's), biologic drugs, and small molecule inhibitors in treating rheumatoid arthritis. However, approximately 30% of patients respond to first-line DMARDs(16), and a further 40% of patients fail to respond to the first choice of biological therapy. Moreover, the diagnosis of rheumatoid arthritis revolves around the exclusion of other diseases that may cause the symptoms. Therefore, identification of predictors that allow accurate diagnosis and prediction of therapeutic outcome would allow the tailoring of treatment, the very definition of precision medicine. Figure 1.5 highlights the number of publications associated with rheumatoid arthritis, which are predominantly associated with treatment, with a small but increasing section associated with prognosis, highlighting this need to personalise treatment.

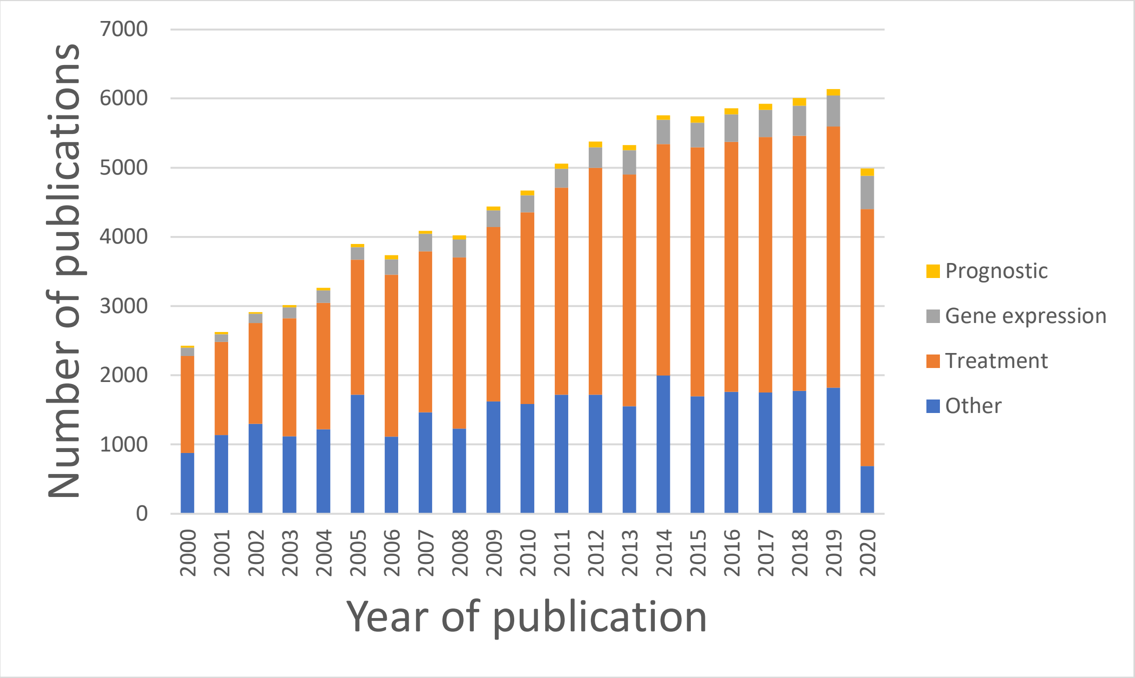


Figure 1.4: Number of publications per year with the term "Rheumatoid Arthritis".

This was also expanded to identify papers that also included other terms; "Prognostic", "Gene expression", and "Treatment".

1.3. Rheumatoid arthritis.

Rheumatoid arthritis is a chronic, systemic autoimmune disease that primarily affects the synovial membranes of the diarthrodial joints. Here, synovial joint inflammation (termed synovitis) is associated with leukocyte infiltration, pannus formation (synovial hyperplasia) and associated damage to bone and cartilage. Untreated, this inflammation leads to progressive and permanent joint damage, and therefore patient disability. Systemically, rheumatoid arthritis can affect induce inflammation in the lungs, pericardium, and skin, and has been associated with disruptions to sleep, cardiovascular disease, anaemia, fatigue, and depression(17). It is heterogeneous in terms of clinical presentation, disease progression, therapeutic response, and tissue pathology. Aspects of this heterogeneity are explored in this introduction, and later in the thesis.

1.3.1. Risk factors in rheumatoid arthritis.

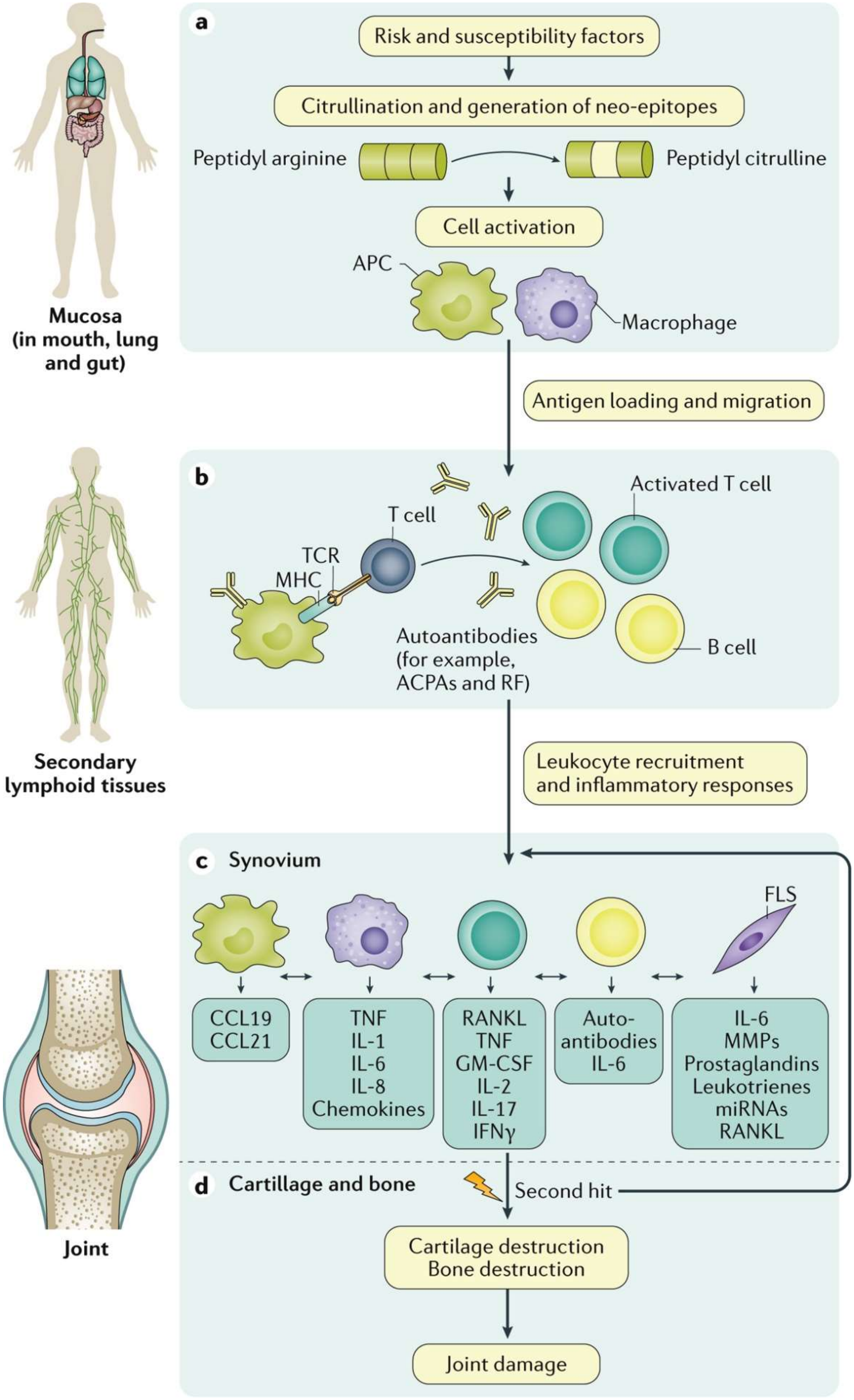
Despite the exact cause of rheumatoid arthritis being unknown(18–20), there is considerable evidence that disease development is associated with several genetic variants(21). 80% of rheumatoid arthritis patients are known to carry *HLA-DRB1* (20), and this in combination with polymorphisms in *PTPN22* it is estimated to account for 40% of the total genetic risk for rheumatoid arthritis (19). Smoking and infection are known environmental factors that can influence development, progression and severity of rheumatoid arthritis(22,23). Additionally, sex is also a risk factor, with women being two-to-three fold more likely to develop the disease(24), which is attributed in part to exposure to hormonal factors such as oestrogen(25).

1.3.2. Pathogenesis of rheumatoid arthritis.

The discovery of rheumatoid factors (RF) – anti-immunoglobulin G (IgG) antibodies – implicated immune dysregulation in the pathogenesis of the disease(26,27). Additional autoantibodies have been discovered to be associated with disease development, particularly anti-citrullinated proteins antibodies (ACPAs)(28,29), as well as anti-carbamylated protein and anti-acetylated protein antibodies(30,31), that can precede the development of rheumatoid arthritis by years. Presence of these autoantibodies (RF, ACPAs, etc) does not mean that the patient will develop inflammatory joint disease, and as such it is hypothesised that it is triggered by a “second hit” signal – such as Epstein-Barr viral infection(32,33). A timeline of this is illustrated in Figure 1.5.

Whilst the adaptive immune system is heavily implicated in the pathogenesis of rheumatoid arthritis, there is also evidence that stromal cells contribute to the development of the disease. Subsets of the fibroblast-like synoviocytes have been identified invading normal cartilage(34), and more recent single-cell experiments characterised fibroblast subsets that drive disease(35).

Returning to the adaptive immune system, early rheumatoid arthritis is characterised by the infiltration of mononuclear cells – primarily CD4$^+$ T cells and macrophages – leading to synovitis (inflammation of the synovial membrane). This accumulation of both lymphoid (T, B, and NK cells) and myeloid (macrophages, neutrophils, mast cells, and dendritic cells) cells induces abnormal angiogenesis, cellular hyperplasia and alterations in the control of extracellular matrix. Here, single-cell RNA-seq data show that many of these activities are coordinated by the intimal lining of the synovium and cell communication between stromal tissue and inflammatory cells(36). Additionally, the presence of the immune cells results in the release of significant amounts of pro-inflammatory cytokines (e.g., tumour necrosis factor, interleukin-1ß, interleukin-6, interferon-γ), chemotactic cytokines (chemokines), growth factors and other inflammatory mediators which control the turnover of extracellular matrix and osteoclastogenesis(37–39).



Nature Reviews | Disease Primers

Figure 1.5: Pathogenesis of rheumatoid arthritis.

A *Genetic risk factors and neo-epitope generation that predisposes patient towards the development of rheumatoid arthritis.* ***B*** *Altered peptides induce autoantibodies.* ***C*** *Stromal cells, Antigen-presenting Cells and macrophages can be activated locally and produce pro-inflammatory cytokines.* ***D*** *Autoimmune response that may require a second hit (e.g. Epstein-Barr virus) inducing joint damage. anti-citrullinated protein antibodies (ACPA), Rheumatoid Factor (RF), major histocompatibility complex (MHC), T cell Receptor (TCR), CC-chemokine ligand (CCL), tumour necrosis factor (TNF), Interleukin (IL), receptor activator of nuclear factor-κB ligand (RANKL), granulocyte-macrophage colony-stimulating factor (GM-CSF), Interferon-γ (IFNγ), Matrix metalloproteinase (MMP), microRNA (miRNA). Figure obtained from Smolen, J. S. et al. (2018) Rheumatoid arthritis Nat. Rev. Dis. Primers doi:10.1038/nrdp.2018.1*

1.3.3. Diagnosis and measuring disease activity.

Whilst the clinical presentation of rheumatoid arthritis is relatively slow, it does develop progressively – accompanied by bouts of aggressive flairs that cause joint damage and reduction to musculoskeletal function(40). Therefore, clinical assessments of disease activity are essential for identifying the stage of the disease and how the patient responds to therapy. Classification criteria are not completely defined, requiring at least one clinically swollen joint that cannot be explained by another disease. A positive diagnosis of rheumatoid arthritis requires a score of ≥6 using the ACR/EULAR 2010 classification criteria as outlined in Table 1.1.

Table 1.1: The 2010 ACR/EULAR classification criteria for rheumatoid arthritis.

For individuals with ≥1 clinically swollen joint not explained by another disease. Large joint defined as one of the following: shoulder, elbow, hip, knee, or ankle. A small joint is defined as the metacarpophalangeal joint, the proximal interphalangeal joint, the second to fifth metatarsophalangeal joints, the interphalangeal joint of the thumb and the wrist. Rheumatoid Factor (RF), Anti-Citrullinated Protein Antibody (ACPA), C-Reactive Protein (CRP), Erythrocyte Sedimentation Rate (ESR).

Joint involvement	Points
1 large joint	0
2-10 large joints	1
1-3 small joints	2
4–10 small joints	3
>10 joints (of which ≥1 is a small joint)	5
Symptom Duration	
<6 weeks	0
≥6 weeks	1
Serology	
RF and ACPA negative	0
Low positive RF or ACPA	2
High positive RF or ACPA	3
Acute-phase reactants	
Normal CRP and ESR	0
Abnormal CRP or ESR	1

Tracking disease progression makes use of numerous disease activity measures, a 2019 review identified 46 methodologies involved in the literature(41). In a clinical setting, only 11 samples met the minimum standard, with 5 methods being preferred for regular use: The Disease Activity Score (DAS) in 28 joints with Erythrocyte Sedimentation Rate (ESR) or C-Reactive Protein (CRP), Clinical Disease Activity Index (CDAI), Simplified Disease Activity Index, Routine Assessment of Patient index 3 (RAPID3), and Patient Activity Scale-II.

Disease Activity Score 28: Utilises a weighted sum of the number of swollen and tender joints (out of 28) in conjunction with ESR or CRP and general health(20,42). Table 1.2 outlines the rules that define the state of disease (high, moderate, and low) as well as the rules associated with tracking response to therapy(43). The weightings are shown in the equation below

$$DAS28 = 0.56\sqrt{TJC +}\,0.28\sqrt{SJC} + 0.7[\ln(esr)] + 0.014(general\ health)$$

Table 1.2: DAS28 scoring and monitoring progression.

Current DAS28		DAS28 decrease from the initial value		
		> 1.2	> 0.6 ≤ 1.2	≤ 0.6
Low	≤ 3.2	Good	Moderate	None
Moderate	> 3.2 ≤ 5.1	Moderate	Moderate	None
High	> 5.1	Moderate	None	None

Clinical Disease Activity Index: the CDAI is based on a simpler composite index, with no weighting on the components. It counts the number of tender and swollen joints (out of 28), patient global assessment of disease activity (0-10 scale), clinical provider global assessment of disease activity (0-10 scale)(44).

Simplified Disease Activity Index: This follows the same concept as the CDAI but includes the levels of CRP in mg/dL (0-10 scale) (45).

Routine Assessment of Patient index 3: RAPID3 is questionnaire-based and is a similar composite index broken into 3 components(46). A questionnaire for activity in the past week (total 0-10 score) followed by a visual analogue scale (VAS) for pain and then for general health.

Patient Activity Scale-II: utilises a weighted sum of Health Assessment Questionnaire Disability Index II (HAQ-II) and VAS for pain and patient global assessment, the weighting is shown below(47).

$$PAS\ II = \frac{(HAQII * 3.33) + Pain + Patient\ Global}{3}$$

Interpretation of the scores for these measures can be seen in Table 1.3.

Table 1.3: Scoring methodologies for Disease activity measures.

CDAI	SDAI	PAS-II	RAPID3	Disease activity
0.0-2.8	0.0-3.3	≤0.25	0-1	Near Remission
2.9-10.0	3.4-11.0	2.6-3.7	1.3-2	Low
10.1-22.0	11.1-26	3.71-8	2.3-4	Moderate
22.1-76.0	26.1-86	>8	4.3-10	High

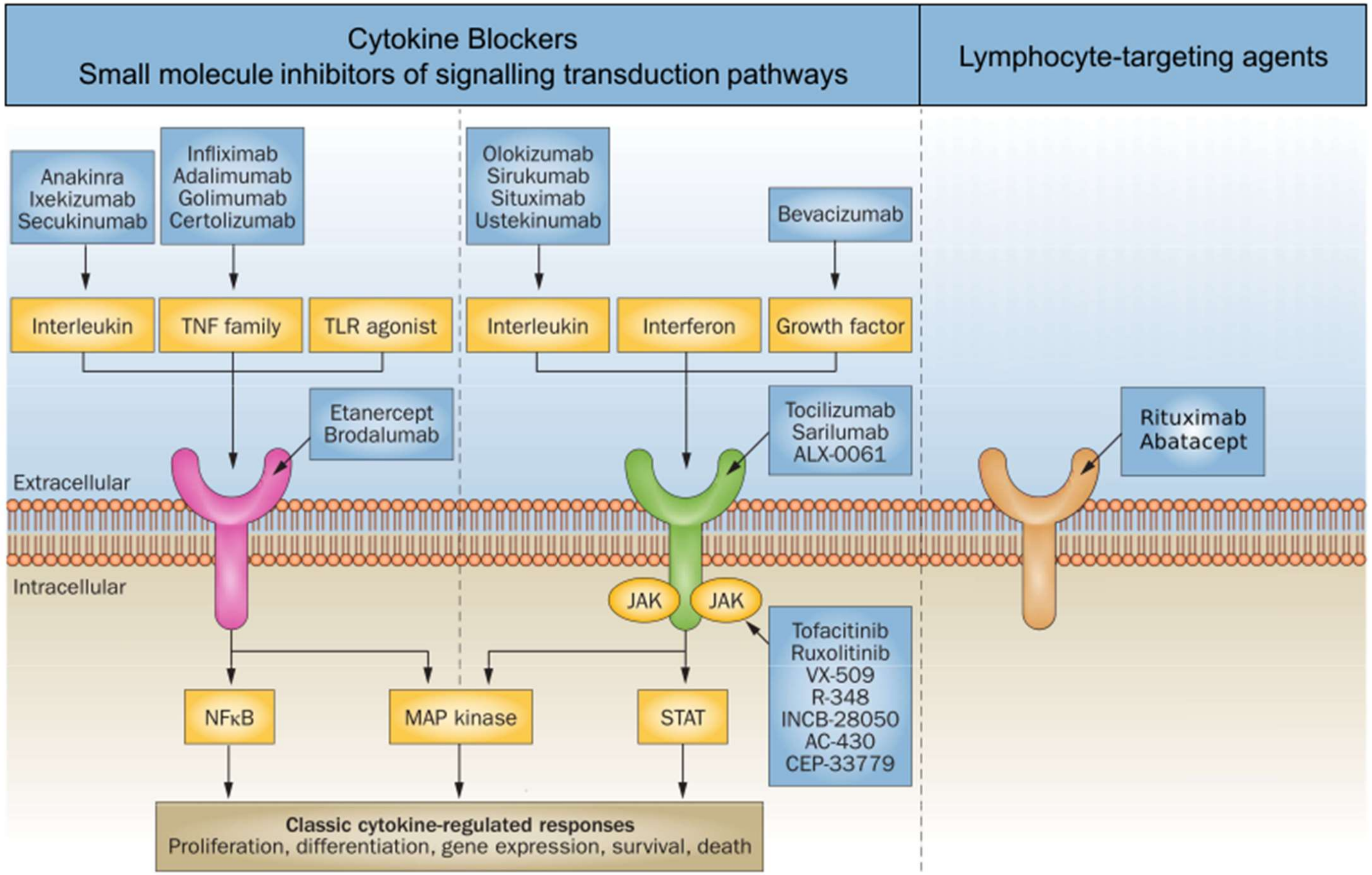


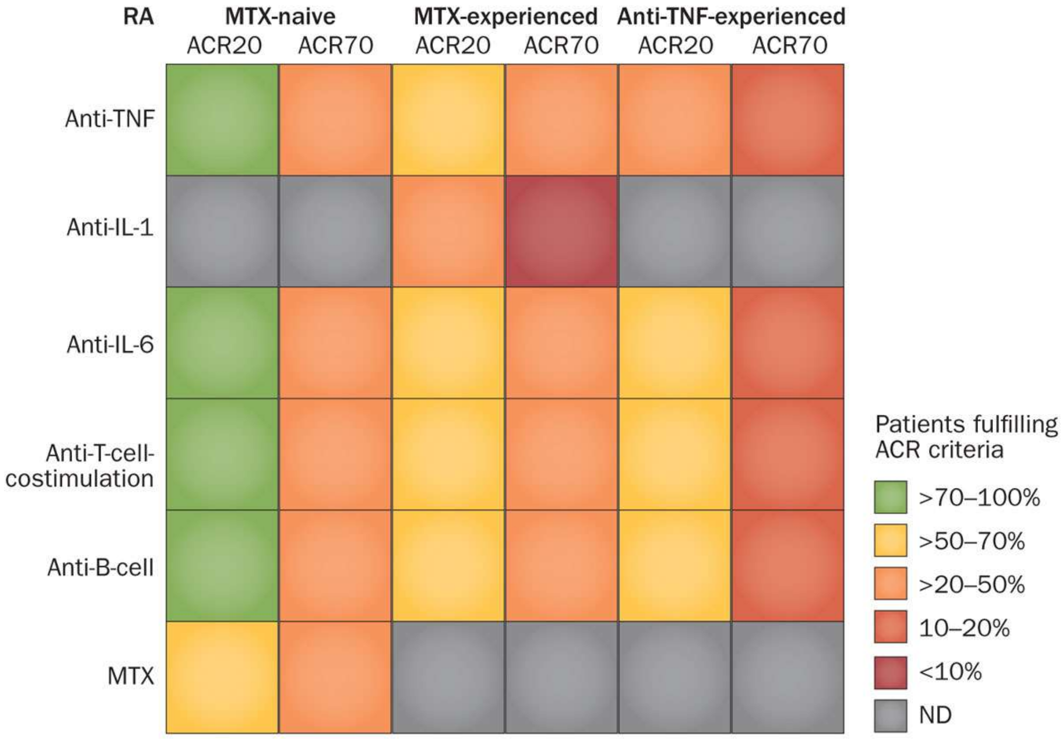
Figure 1.6: Biologics and small-molecule inhibitors used in the treatment of rheumatoid arthritis.

Treatments target cytokines (eg Interleukin, Tumour Necrosis Factor (TNF), Interferons) or their Receptors and downstream signalling pathways. Additionally, cell depletion (Rituximab) or suppressing co-stimulation (Abatacept) provide further pathways to target. Toll-Like Receptor (TLR), Nuclear Factor KB (NFKB), Janus-Activated Kinase (JAK), Mitogen-Activated Protein (MAP), Signal Transducer and Activator of Transcription (STAT). Figure adapted from Choy, E.H., Kavanaugh, A.F., & Jones, S.A. (2013) The problem of choice: current biologic agents and future prospects in RA. Nature Reviews Rheumatology.

1.3.4. Therapeutic intervention in rheumatoid arthritis.

Treatment of rheumatoid arthritis revolves around reducing systemic and local inflammation, and thereby prevent irreversible joint damage from occurring. Early diagnosis and treatment is key to an effective therapeutic response(48), and treatment guidelines follow a prescribed regimen of treatments utilising a treat-to-target approach with disease-modifying anti-rheumatic drugs (DMARD's), biological drugs, or small molecule inhibitors(16). Figure 1.6 illustrates the selection of biological drugs and small molecule inhibitors, as well as the pathways targeted.

Upon diagnosis, immediate treatment using DMARD's such as methotrexate are recommended(49). If disease activity is not controlled by the therapeutics after 3-6 months, methotrexate may be supplemented with TNF inhibitors (infliximab, adalimumab, etanercept) unless contraindicated. The National Institute for Health and Care Excellence (NICE) guidelines(50) that illustrate this prescribed regimen are shown in Figure 5.1.



Nature Reviews | Rheumatology

Figure 1.7: Therapeutic responses for patients at different stages of disease and treatment exposure.

Data is broken into treatment naïve (early RA), experienced to methotrexate (MTX) and experienced to anti-TNF therapy. Anakinra (anti IL-1) Tocilizumab (anti IL-6), abatacept (anti T-cell costimulation), Rituximab (anti-B-cell). Cells are coloured to demonstrate the proportion of patients who respond to treatment. Figure obtained from Smolen, J. & Aletaha, D. (2015 Rheumatoid arthritis therapy reappraisal: strategies, opportunities and challenges. Nat Rev Rheumatol

One important fact to consider, however, is that only 30% of patients respond to methotrexate monotherapy(51), and 40% of patients will have a poor response to anti-TNF therapy(16,52). This variability in response is one of the driving concepts behind the need for precision medicine in rheumatology, maximising the therapeutic window with the aim to increase the chances of drug-free remission(53,54). Figure 1.7 shows a matrix of response rates to therapeutics for patients through the course of disease and exposure to therapy.

First-line DMARDs attempt to control the disease through suppression of inflammation, whilst more targeted therapies generally operate by blocking cytokine signalling or by targeting lymphocytes.

1.3.4.1. Blocking Cytokine signalling.

Anti-TNF: Five biologic agents are approved for therapeutic targeting of TNF in rheumatoid arthritis, with numerous biosimilars now entering the market(55). The five agents are the monoclonal antibodies; infliximab, adalimumab, certolizumab pegol, and golimumab, and the fusion protein etanercept(56).

Anti-IL-6: Tocilizumab is a monoclonal antibody that acts as an IL-6 receptor agonist. It, therefore, targets both canonical signalling through membrane-bound IL-6R as well as trans-signalling through soluble IL-6R(57,58).

Anti-Il-1: Anakinra is a recombinant form of human IL-1ra, whilst it has shown to be effective, the absolute differences in ACR23, ACR50 and ACR70 is lower than treatments such as etanercept or adalimumab(59).

Small molecule inhibitors: Tofacitinib is a small molecule inhibitor that targets JAK1 and JAK3, thereby interfering the Jak-STAT pathway that is downstream on many cytokine signalling evens (Figure 1:6)(60).

1.3.4.2. Targeting lymphocytes.

Anti-T-cell co-stimulation: Abatacept is a fusion protein of *CTLA4* and the FC domain of IgG1. CTLA4 preferentially binds CD80/CD86 and by doing inhibits the transmission of the costimulatory signal from antigen-presenting cells, and thereby T-cell activation(61)

Anti-B-Cell: Rituximab is a monoclonal antibody targeting CD20 and in so doing results in the depletion of B-cells(62).

1.3.4.3. Adverse effects.

With the suppression of inflammation and other components of the immune system, this results in issues associated with opportunistic infections. Moreover, these therapeutic interventions are associated with their own adverse effects(63).

The adverse effects of methotrexate are generally associated with the dosage, low dosage side effects are generally mild: haematological abnormalities, gastrointestinal problems, the elevation of liver enzymes. But ultimately fewer than 5% need to discontinue due to adverse effects(63)

Anti-TNF therapies often result in reactions around the site of injection – itching, pain, and redness. However, these therapies are also associated with reactivation of tuberculosis, demyelinating diseases and skin cancer(64).

1.3.4.4. Trial and error.

Despite numerous therapies available, significant proportions of rheumatoid arthritis patients fail to respond to treatment. Moreover, as is demonstrated in Figure 1.7, the more the disease progresses and is exposed to therapy the lesser the chance of good response and achievement of remission. As outlined in Figure 5.1, treatment follows a proscribed regimen of therapies, waiting for clinical response after 3-6 months. Not only is this approach inefficient, but it also leaves open the potential for irreversible damage to occur during these periods of uncontrolled disease. Moreover, from a financial perspective, this results in the wasteful usage of valuable therapeutics for limited if any benefit.

This highlights the need to tailor the therapeutic intervention to the individual, necessitating a method to predict clinical response. The identification of the correct treatment during the window of opportunity early in the disease is essential in maximising the opportunity for drug-free remission(48). Additionally, this avoids excessive exposure to unnecessary therapeutics, and the adverse effects associated therein.

1.3.5. Synovitis pathologies.

Histopathological analysis of synovitis defines the characteristic features as hypertrophy of the lining layer, neo-angiogenesis, and infiltration of immune cells. Immune cells are observed as two broad patterns that can overlap – randomly distributed throughout the sub-lining, or organised into follicular structures - defined as diffuse and follicular respectively(65–69). Moreover, there is a third pattern with relatively little immune infiltrate, termed pauci-immune that still represents active disease, supporting the role of stromal cells in driving synovitis. Immunophenotypic characterisation of these pathologies is illustrated in Figure 1.8, staining for lymphocytes (T, B, and plasma cells) and macrophages(65,67,70,71).

- Follicular synovitis is characterised by enrichment for lymphoid infiltrate (particularly B-cells) that may be arranged in aggregates called ectopic lymphoid structures that function as germinal centres.
- Diffuse synovitis is characterised by a primarily myeloid infiltrate, with little B cell presence
- Pauci-immune demonstrates minimal immune infiltrate.

This heterogeneity in tissue pathology has been associated with the therapeutic outcomes, for example, pauci-immune patients not responding to TNF inhibition(72), whilst patients with the diffuse pathology were most likely to benefit from TNF inhibition(73)

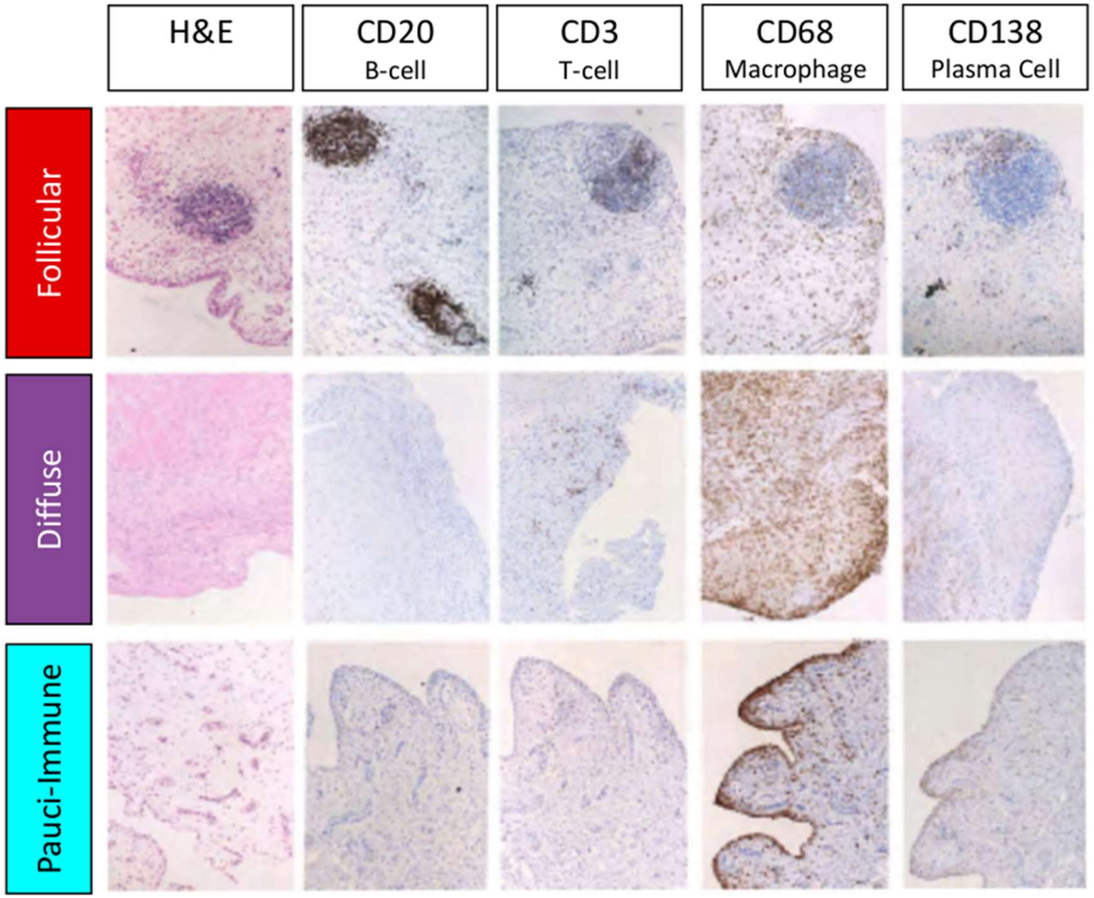


Figure 1.8: Immunophenotyping of the synovitis pathologies.

The three patterns of synovitis seen in the pathologies of rheumatoid arthritis. Follicular exhibits strong staining of lymphoid cells, and the presence of ectopic lymphoid structures. Diffuse exhibits high levels of macrophage staining throughout the tissue, with some T cells, but a scarcity of B cells. Pauci-immune shows almost no immune infiltrate, but high levels of macrophage staining in the sub-lining layers of the synovial membrane. Figure adapted from Pitzalis C, Kelly S, Humby F (2013) New learnings on the pathophysiology of RA from synovial biopsies. Curr Opin Rheumatol 25, 334–44.

1.3.6. Predictive signatures in rheumatoid arthritis.

Over the years, numerous signatures and rule sets have been developed aiming to improve the understanding of rheumatoid arthritis utilising synovial biopsies, cell cultures, and peripheral blood mononuclear cells (PBMC)(74). Given the window of opportunity for drug-free remission necessitates early treatment to prevent irreversible joint damage from occurring. Because of this, the EULAR guidelines recommend the measurement of ACPA's in early diagnosis(75), however, this introduces even more heterogeneity as only 50% of patients test positive(76). Identification of biomarkers is therefore a topic of intensive interest.

A systemic review in 2016 identified 57 studies that utilised 79 (bio)markers and 8 multivariable models that resulted in 14 predictors(77). These utilised EULAR, ACR20/50, and DAS28 response criteria to assess how well predictor of erosive disease, therapeutic response after methotrexate usage, environment risks (smoking, Epstein-Barr Virus exposure), presence of RF (and immunoglobulin specific forms) or ACPAs, and genotypic risks.

Further studies have identified gene signatures that discriminate rheumatoid arthritis from non-rheumatoid arthritis. In 2004, microarray analysis resulted in a set of 63 genes that discriminated between rheumatoid arthritis and osteoarthritis(78), further work identified a 12 gene signature that did the same(79). Yet despite these have not been comprehensively validated in other datasets, and therefore are not utilised in clinical practice.

Clinical response to therapeutics has been a major focus of these predictive studies, numerous studies(71–73,80,81,81–94) aimed to identify those features that will allow stratification of response. Exploration of the gene expression has been extensively used to attempt to identify predictive signatures.

Over the years, many genes have been identified as discriminating responders and non-responders for multiple therapies:

- **Methotrexate**: Two studies identified genes – 133 and 16 genes respectively – that differentiated responder, these studies were limited to conference proceedings, and have not been validated in external cohorts(95,96). Due to the limited information in these proceedings, it is also impossible to see how these signatures overlap.
- **Infliximab**: Multiple studies have identified several genes(82,97–101). Whilst none of these genes are found in all the studies, however, several genes were found in multiple studies such as *CCL19, HLA-DQA1, IL2RB* and *FCGR1A*.

- **Rituximab**: Studies have identified multiple signatures(102,103) that discriminate responders, however, these show no overlap between signatures. However, it does reveal several genes that overlap with infliximab, such as *HLA-DQA1, MxA,* and *MxB*.
- **Tocilizumab**: One study identified 59 genes that discriminate response(104), but without additional signatures to compare there are no core genes to investigate.

One of the largest challenges with identification of biomarkers is the lack of hard definitions, in oncology, there is survival or remission, whilst in rheumatoid arthritis progression of the disease is more subtle. Adding to this, many of these biomarkers are never replicated in future cohorts. For example, whilst serum levels of calprotectin (*S100A8/A9)* has been associated in multiple studies(105,106) with disease activity in rheumatoid arthritis and response to anti-TNF therapy, large cohort studies failed to replicate this(92). Further challenges result from most studies being small, without sufficient sample sizes, these signatures are prone to overfitting.

1.4. Aims.

Precision medicine requires the identification of biomarkers that are capable of discriminating subclasses of disease, be that pathology, therapeutic outcome, or survival. To implement this clinically these markers need to be robust and capable of being validation in external datasets and the clinic. Rheumatoid arthritis is a complex autoimmune disease with distinct subclasses that are associated with therapeutic outcome and therefore provides an ideal candidate to interrogate for biomarkers.

The main goals of this thesis are:

- **Stratification of synovitis pathologies based on transcriptome:** Current methodologies for differentiating the pathologies revolves around the histological assessment of immune infiltrate into the synovitis. Using transcriptomics is it possible to identify the contribution of immune cells, and therefore discriminate the form of synovitis seen in the patient.
- **Define a characteristic transcriptional profile for the pathologies:** Using stratified samples create an archetypical profile that allows comparison of other samples with less clear-cut disease. This may improve the identification of “grey area” patients who lack defined characteristics of synovitis and indicate more appropriate therapies.
- **Pathology specific molecular pathways:** Using stratified transcriptomes, differential gene expression and downstream pathway analysis may provide insights into the pathogenesis of the disease. This also offers the opportunity to identify novel

molecular targets that are more specific to the pathology, allowing the potential of targeted therapies that avoid unwanted systemic effects.

- **Identification of Biomarkers:** Based on stratified transcriptional profiles, is it possible to identify biomarkers – in this case, genes - that offer the potential to use a classifier of disease. Moreover, these biomarkers need to show themselves to be robust, and replicable across multiple clinical cohorts.

2. Materials and Methods.

2.1. Reagents.

Unless otherwise stated, all reagents were purchased from Sigma-Aldrich.

2.2. In vivo experiments.

2.2.1. Mice strains.

All animal work was performed in accordance with the United Kingdom Animals (Scientific Procedures Act 1986), and under the authority of the Home Office Personal (PIL: IBBC24E9D) and Project (PPL: PB3E4EE13) Licences.

8 to 12 week-old wild type mice (C57Bl/6) were purchased from Charles River. Interleukin-6 receptor (*Il6ra*$^{-/-}$, or CD126$^{-/-}$) and interleukin-27 receptor subunit alpha (*Il27ra*$^{-/-}$, or *Wsx1*) deficient mice on a C57Bl/6 background were bred in house (PPL: PCIFFFEE3 & PB3E4EE13 respectively).

Il6ra$^{-/-}$ mice were generated at GlaxoSmithKline (Stevenage, U.K.) by disrupting exons 4,5, and 6 by insertion of a neomycin cassette through recombinase-mediated cassette exchange (107).

Il27ra$^{-/-}$ mice were originally sourced from The Jackson Laboratory (line B6N.129P2-*Il27ra*^{tm1Mak}/J) and were generated in the same manner with a neomycin cassette disrupting the fibronectin type III domain of *Il27r* (108).

2.2.2. Antigen-Induced Arthritis.

Antigen-induced arthritis (AIA) is a monoarticular model of inflammatory arthritis induced by an intra-articular administration of antigen into the joint of antigen primed mice (109–112). This leads to synovitis and alterations in both cartilage and bone remodelling resembling clinical inflammatory rheumatoid arthritis. The timeline for development of disease is illustrated in Figure 2.2.

Mice were primed by immunising against the antigen – methylated Bovine Serum Albumin (mBSA) at two timepoints: day -21 and -14 (relative to arthritis induction) by subcutaneous (s.c.) injection of 100 µl of 1 mg/ml mBSA/Complete Freud's Adjuvant emulsion using a 1ml syringe and 25G needle. Antigenic response was modulated with the administration of 100 µl of 1.6µg/ml heat-inactivated pertussis toxin intraperitoneally (i.p.) (1ml syringe, 25g needle).

mBSA/CFA emulsion was prepared by dissolving 10 mg mBSA (A1009, Sigma Aldrich) in 5ml sterile water (PL 1502/003R) and combining with 5ml CFA (F5881, Sigma Aldrich). The mBSA/CFA mixture was passed through a 18G needle ~ 20 times until a stable, white emulsion which forms a defined sphere that doesn't easily disperse when dropped into water. 8µl heat-inactivated Pertussis Toxin (P2980, Sigma Aldrich) in 1ml sterile water to obtain a final concentration of 1.6µg/ml.

Arthritis was induced by injecting antigen with a 10 µl intra-articular (i.a.) injection of mBSA (10 mg/ml) into the hind limbs using a 29G insulin needle.

Following induction mice were sacrificed by schedule 1 method (CO_2 followed by cervical dislocation) at two timepoints: an early stage (day 3) – representing an acute inflammatory synovitis, and a late stage (day 10) – representing chronic inflammation. Additionally, naïve mice were sacrificed to provide non-inflamed synovium that is used as a baseline. The Synovial tissue collected was stored in RNAlater (AM7024, ThermoFisher Scientific) at -80°C.

Animal wellbeing was monitored with regular inspections (minimum 3x a week) with the animals being weighed when carrying out regulated procedures. Following arthritis induction or adverse effects from CFA (ulceration) monitoring is increased to daily inspections. In the case of ulceration, mice are treated with topical application of iodine (3030440, Farla Medical) and weighed. Arthritis severity was assessed by comparing the joint diameter relative to baseline using a POCO 2T micrometer (Krœplin) (109), Figure 2.1 is an demonstrates the change in joint diameter reflecting the development of arthritis.

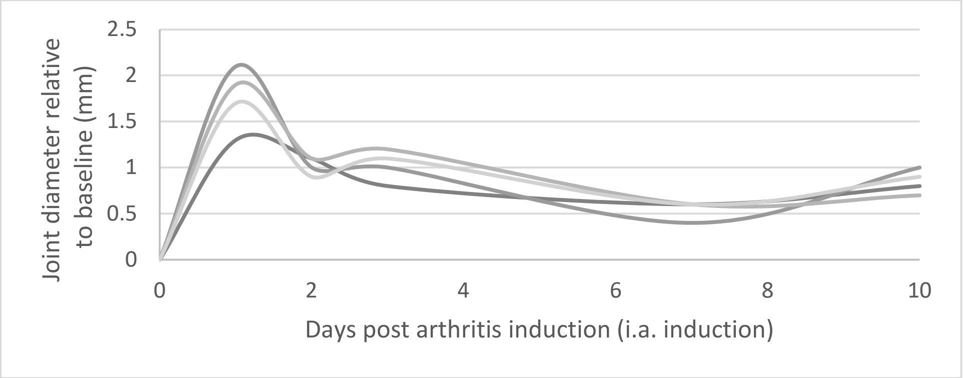


Figure 2.1: Joint diameter measurements of Il6r mice post induction.

As both knees were injected to minimise the number of mice used whilst maximising the tissue harvested, measurements are relative to baseline before AIA is induced. Chromatin Immunoprecipitation.

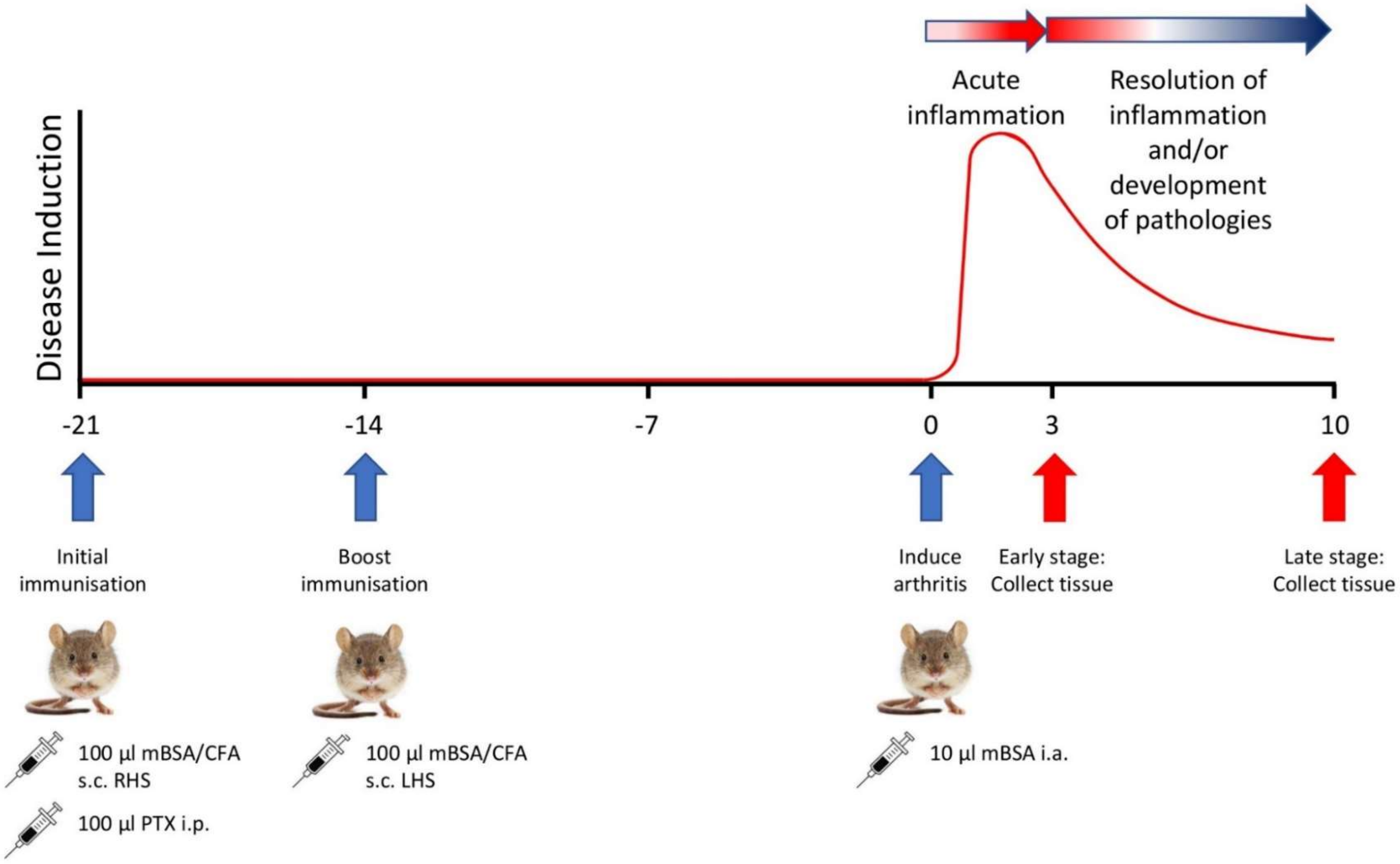


Figure 2.2: Timeline of procedures and the resulting joint swelling and disease stages of antigen induced arthritis.

Blue Arrows illustrate timepoints of regulated procedures, whilst red arrows indicate timepoints at which mice are sacrificed following induction. Initial priming induces the creation of a population of T-cells that are reactive against the mBSA antigen prior to the development of synovitis. Following induction, the acute inflammation is driven by an early T-cell response and infiltration of innate monocyte populations. The chronic-like phase is characterised by prominent T- and B cell infiltrates. s.c., subcutaneous, i.p., intraperitoneal, i.a., intraarticular.

2.3. Chromatin Immunoprecipitation.

Chromatin Immunoprecipitation (ChIP) allows the investigation of protein-DNA interactions (113–115). The basic protocol for ChIP analysis is illustrated in Figure 2.3.

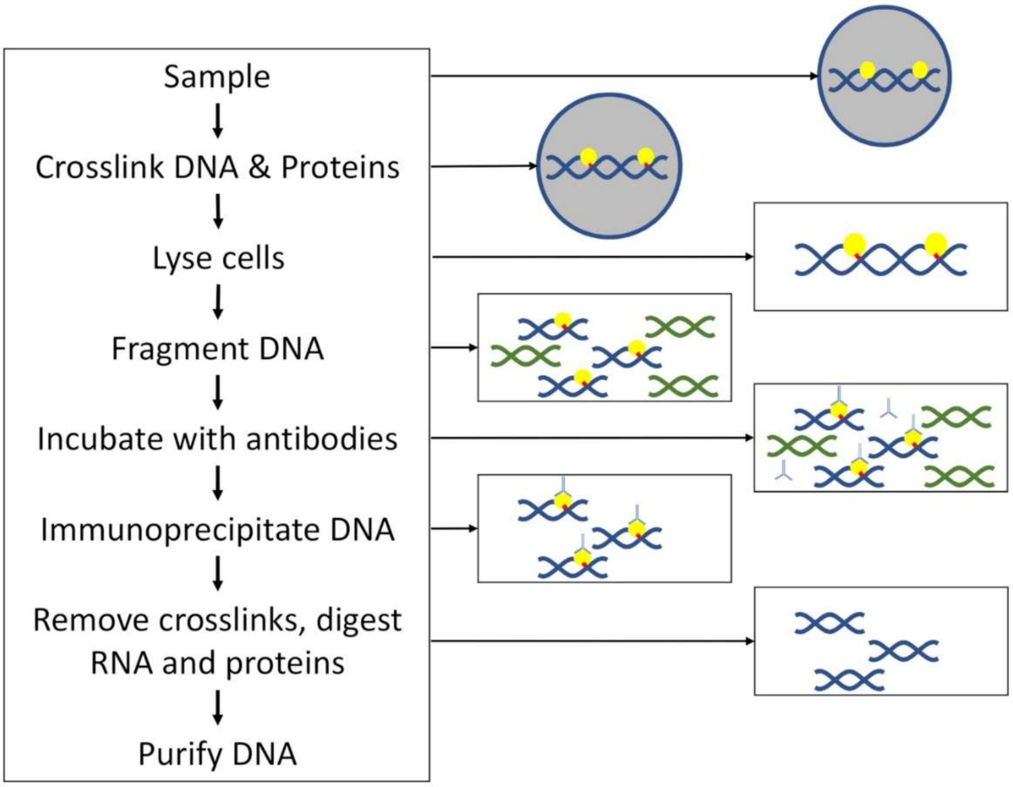


Figure 2.3: Workflow involved in chromatin immunoprecipitation.

DNA and proteins (yellow circles) are crosslinked using formaldehyde (red), before cellular and nuclear lysis. DNA is fragmented before samples are incubated with antibodies; immunoprecipitation is done using protein A/G magnetic beads. Crosslinks are removed, before digesting RNA and proteins, and purifying DNA for further analysis.

In summary (116), proteins are reversibly crosslinked to the DNA using formaldehyde, and the cells and nuclei are lysed. The DNA is then fragmented to 200-400 bp lengths by sonication, and the DNA-protein fragments are incubated with an appropriate antibody, before precipitation using protein A/G magnetic beads . The crosslinked proteins are released and digested, and DNA purified for further investigation.

2.3.1. Adaptations for ChIP on tissue.

The existing laboratory method for ChIP was based on protocols optimised for T-cells analysis. This method was modified for the analysis of synovial tissue and required the introduction of a tissue disruption step to disaggregate the cells from the extracellular tissue architecture. It was however critical to ensure that this adaptation to the protocol did not affect the ability to crosslink protein bound to the DNA. All buffers utilised in this section are detailed in Table 2.2.

Tissue disaggregation: Synovial tissue was removed from the RNAlater and weighed to achieve approximately 10 mg per pulldown (aim for ~> 20 mg tissue for Stat1 and Stat3 pulldowns). The tissue was snap frozen in liquid nitrogen, and ground to a powder using a disposable Axygen tissue grinder (12649595, ThermoFisher Scientific).

Crosslinking DNA and proteins: The ground samples were defrosted in 1ml crosslinking buffer and incubated at room temperature for 15 minutes, cells are then pelleted by spinning at 500xg for 1 minute at room temperature. Crosslinking is halted by resuspending pellet in "Stop solution" and shaking for 1 minute, before pelleting cells (500xg, 1min, at room temperature). Cells were washed with 1ml PBT and spun down (500xg, 1min, at room temperature).

Lyse cells: The cell pellet was resuspended in 1ml "Cell Lysis Buffer" and incubated on ice for 10 minutes, gently agitating every few minutes, before pelleting the nuclei (500xg, 5min, 4°C). Nuclei are the resuspended in 275µl "Nuclear Lysis Buffer" (NLB) and incubated on ice for 10 minutes, before diluting with 165µl "IP Dilution Buffer" (IPDB).

Fragment DNA: DNA was then fragmented using Bioruptor Plus (Diagenode) for 30 cycles (High intensity, 30 seconds on, 30 seconds off, at 4°C). Sample is diluted to obtain a final ratio of 1:4 (NLB:IPDB), with an additional 935 µl "IP Dilution Buffer", (total volume 1375 µl) before separating into aliquots for Input (100 µl), IgG (100 µl) Stat1 (550 µl), Stat3 (550 µl), and for gel (20 µl) to check shearing size.

Shearing of the genomic DNA was checked by removing crosslinks and digesting protein (20 µl sample, 1.4 µl NaCl (5M) 2.6 µl H20, 1 µl Proteinase K (AM2548, ThermoFisher Scientific), incubating at 65°C for 1-2 hours. Samples were labelled with 6x loading dye (R0611, ThermoFisher Scientific) and loaded onto 2% (w:v) agarose gel with SYBR-safe (S33102, ThermoFisher Scientific) with 100bp GeneRuler ladder (SM0242, ThermoFisher Scientific) and running at 100v for 30-45 minutes.

Incubate with antibodies: Samples were incubated with antibody overnight at 4°C on a rotating wheel, Isotype IgG aliquots from samples at each time point was pooled. Stat1 (#9172) used at 1:100, Stat3 (C20) used at 4 µg (20 µl), IgG (C15410206) used at 4 µg (4 µl). Details of antibodies used are listed in Table 2.1. Isotype IgG acts as a control for non-specific binding.

Immunoprecipitate DNA: Magnetic protein A/G beads (78609, ThermoFisher Scientific) are thoroughly vortexed to ensure complete suspension, beads were washed in cold PBS (40 µl beads per pulldown) using magnetic stand to pellet the beads before resuspending in samples and incubating for 2 hours at room temperature on a rotating wheel.

Samples then go through a series of washes, all of which utilised 1 ml buffer on a rotating wheel for 5 minutes at 4°C; twice with "IP Wash Buffer 1", twice with "IP Wash Buffer 2", and twice with TE. The samples were then resuspended in 100 µl TE and transferred to a LoBind Eppendorf. Samples and input aliquots have of 10 µl 10% SDS, 6 µl 5M NaCl and 2 ng RNaseA (R6513) added before incubating at 65°C for 2-4 hours. Sample supernatant is then transferred to a LoBind Eppendorf, and to maximise immunoprecipitated DNA the beads are washed with 100 µl TE and supernatants combined, an additional 6 µl 5M NaCl and 5 µl Proteinase K before incubating overnight at 45 °C.

DNA purification: DNA is purified by combining the sample with an 200 µl phenol:chloroform:isoamyl alcohol (25:24:1) (P2069-100ML) vortexing thoroughly before centrifuging at 16'000g for 5 minutes at room temperature. The aqueous layer containing DNA, is transferred to a LoBind Eppendorf with 20 µl 3M sodium acetate (pH 5.2) (neutralising phosphate backbone charge leading to precipitation) and 10 µg glycogen (which acts as a carrier by chelating small nucleotide fragments) before thoroughly vortexing, before adding 500 µl 100% ethanol. DNA is then precipitated at -80°C for 1-2 hours, then centrifuging at 16800g for 20 minutes at 4°C. The DNA pellet is then washed in ice cold 70% ethanol before spinning down again (16800g, 10min, 4°C), before leaving the pellet to air dry for ~15 minutes. The DNA pellet is resuspended in 30 µl ultrapure water

2.3.2. Antibodies.

Table 2.1 outlines the antibodies utilised for ChIP.

Table 2.1: Antibodies utilised for chromatin immunoprecipitation.

All precipitations utilised an isotype control for non-specific binding.

Target	Source	Species Reactivity	Clone	Isotype	Type	Company
STAT1	Rabbit	Human, Mouse, Rat	m-22	IgG	Polyclonal	Santa Cruz
STAT1	Mouse	Human, Mouse, Rat	C-136	IgG1	Monoclonal	Santa Cruz
STAT1	Rabbit	Human, Mouse, Rat, Monkey	D1K9Y	IgG	Monoclonal	Cell Signalling Technologies
STAT1	Rabbit	Human, Mouse, Rat, Monkey	#9172	IgG	Polyclonal	Cell Signalling Technologies
STAT3	Rabbit	Human, Mouse, Rat, Xenopus	C-20	IgG	Polyclonal	Santa Cruz
Non-specific	Mouse	-	X0931	IgG1	Monoclonal	Dako
Non-specific	Rabbit	-	C15410206	IgG	Polyclonal	Diagenode

2.3.3. Buffer composition.

The composition of buffers used in the extraction and preparation of samples for chromatin immunoprecipitation are listed in Table 2.2.

Table 2.2: Buffers utilised in chromatin immunoprecipitation.

	Compound	Final Concentration
Crosslinking Solution	Ethylene glycol-bis(2-aminoethylether)-N,N,N′,N′-tetraacetic acid (**EGTA**) pH 8	0.5 mM
	Ethylenediaminetetraacetic acid (**EDTA**) pH 8	1 mM
	2-[4-(2-hydroxyethyl)piperazin-1-yl]ethanesulfonic acid (**HEPES**)	50 mM
	NaCl	100 mM
	Formaldehyde	1.5% (v/v)
Stop Solution	Glycine	125 mM
	Triton X-100	0.1% (v/v)
	Phosphate Buffered Saline (**PBS**)	-
PBT	Triton X-100	0.1% (v/v)
	PBS	-
Cell Lysis Buffer	Tris-HCL pH 8.1	10 mM
	NaCl	10 mM
	Nonyl phenoxypolyethoxylethanol (**NP-40**)	0.2% (v/v)
	Sodium butyrate	10 mM
	phenylmethanesulfonyl fluoride (**PMSF**)	50 µg/ml
	Leupeptin	1 µg/ml
Nuclear Lysis Buffer	Tris-HCL pH 8.1	50 mM
	EDTA pH 8	10 mM
	Sodium dodecyl sulfate (SDS)	1% (w/v)
	Sodium butyrate	10 mM
	PMSF	50 µg/ml
	Leupeptin	1 µg/ml
IP Dilution Buffer	Tris-HCL pH 8.1	20 mM
	NaCl	150 mM
	EDTA pH 8	2 mM
	Triton X-100	1% (v/v)
	SDS	0.01% (w/v)
	Sodium butyrate	10 mM
	PMSF	50 µg/ml
	Leupeptin	1 µg/ml

IP Wash Buffer 1	Tris-HCL pH 8.1	20 mM
	NaCl	50 mM
	EDTA pH 8	2 mM
	Triton X-100	1% (v/v)
	SDS	0.01% (w/v)

IP Wash Buffer 2	Tris-HCL pH 8.1	10 mM
	LiCl	250 mM
	EDTA pH 8	1mM
	Triton X-100	1% (v/v)
	SDS	0.01% (w/v)

TE	Tris-HCL pH 8.1	10 mM
	EDTA pH 8	1mM

2.4. ChIP-qPCR.

Quantitative Polymerase Chain Reaction (qPCR) of known STAT associated genes determines the levels of enrichment achieved in the immunoprecipitation (117).

Oligonucleotide primers were designed with coverage of promoter sequences annotated with STAT binding sites, as well as negative controls downstream of the gene. Primers were designed by downloading the promoter sequences of known STAT associated genes, these were analysed for Gamma interferon activation site (GAS) motifs. Primers were designed to cover these GAS sites, and STAT binding was checked in the better annotated human genome by blasting the PCR sequences. Negative control sequences were created similarly by looking downstream of the transcription start site, but with the intention of identifying sequences with no GAS motifs. Downstream sequences were also checked in human genome, again because of the better annotation. Figure 2.3.2 illustrates the process using the known Stat1 binding gene *Irf1*.

All PCR runs were carried out using 10 µl reaction volume, with 1 µl of sample and 9 µl of primer master mix ((5 µl TaqMan Fast Advanced Master Mix (4444557, ThermoFisher Scientific), 4.5 µl nucleotide free water, 0.5 µl primer) mix per reaction). Samples were run on either the ViiA7 or QuantStudio 12k Flex Real-Time PCR systems for 96 or 384well plates respectively. Samples were run in triplicate for each primer used, in addition to a negative well for each primer used. Samples were run for 45 cycles using the standard TaqMan Fast protocol, as listed in Table 2.3.

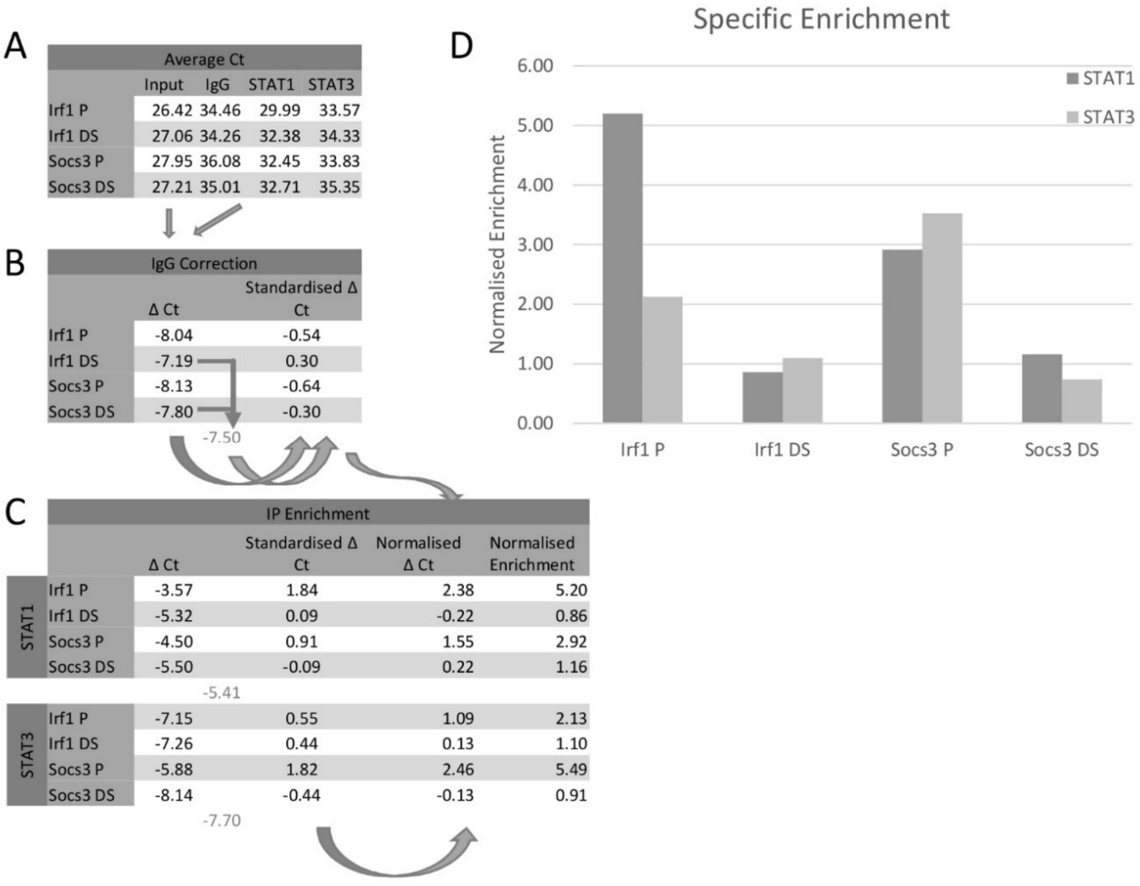
Table 2.3: TaqMan Fast protocol used for ChIP-qPCR.

	Stage	Temperature	Time (seconds)
Uracil-N-Glycosylase incubation	Hold	50°C	120
Polymerase activation	Hold	95°C	20
PCR (45 cycles)	Denature	95°C	1
	Anneal/Extend	60°C	20

2.4.1. Calculation of specific enrichment from ChIP-qPCR results.

To determine whether the immunoprecipitation was successful, we look at enrichment of DNA fragments from our positive controls compared negative regions and IgG isotype controls using qPCR. To calculate this enrichment, we use the following steps, also illustrated in Figure 2.4:

- IgG ΔCt was determined by subtracting the average Input Ct from average IgG Input (A).
- This was then normalised by subtracting the average ΔCt of the negative controls (primers for the downstream sequences) (B)
- IP ΔCt was calculated in the same manner.
- IP normalised ΔCt was then calculated by subtracting the IgG ΔCt from the IP ΔCt (C).
- IP normalised enrichment values were then expressed as 2^$^{\text{IP normalised } \Delta\text{Ct}}$ (C & D).



A

Average Ct	Input	IgG	STAT1	STAT3
Irf1 P	26.42	34.46	29.99	33.57
Irf1 DS	27.06	34.26	32.38	34.33
Socs3 P	27.95	36.08	32.45	33.83
Socs3 DS	27.21	35.01	32.71	35.35

B

IgG Correction	Δ Ct	Standardised Δ Ct
Irf1 P	-8.04	-0.54
Irf1 DS	-7.19	0.30
Socs3 P	-8.13	-0.64
Socs3 DS	-7.80	-0.30
	-7.50	

C

IP Enrichment		Δ Ct	Standardised Δ Ct	Normalised Δ Ct	Normalised Enrichment
STAT1	Irf1 P	-3.57	1.84	2.38	5.20
	Irf1 DS	-5.32	0.09	-0.22	0.86
	Socs3 P	-4.50	0.91	1.55	2.92
	Socs3 DS	-5.50	-0.09	0.22	1.16
		-5.41			
STAT3	Irf1 P	-7.15	0.55	1.09	2.13
	Irf1 DS	-7.26	0.44	0.13	1.10
	Socs3 P	-5.88	1.82	2.46	5.49
	Socs3 DS	-8.14	-0.44	-0.13	0.91
		-7.70			

Figure 2.4: Illustrating the calculation for specific enrichment, arrows show which values are utilised for each step.

***A**; Average Ct values from qPCR using auto-thresholding. **B**; IgG correction calculated ΔCt by subtracting IP Ct (in this case IgG) from Input Ct. this is then standardised by subtracting an average of the negative controls (downstream) (greyed out number). **C**; Calculation of ΔCt and standardised ΔCt is performed in same manner as the IgG correction. Values are then normalised by subtracting IgG standardised ΔCt from IP standardised ΔCt. Normalised enrichment (1 being no enrichment) is calculated as 2^normalised ΔCt. **D**; Plotting the normalised enrichment shows the preferential binding of STAT1 to Irf1 promoter, whereas Socs3 promoter has similar STAT1 and STAT3 binding. Note that both negative controls (downstream) show no enrichment (with values around 1).*

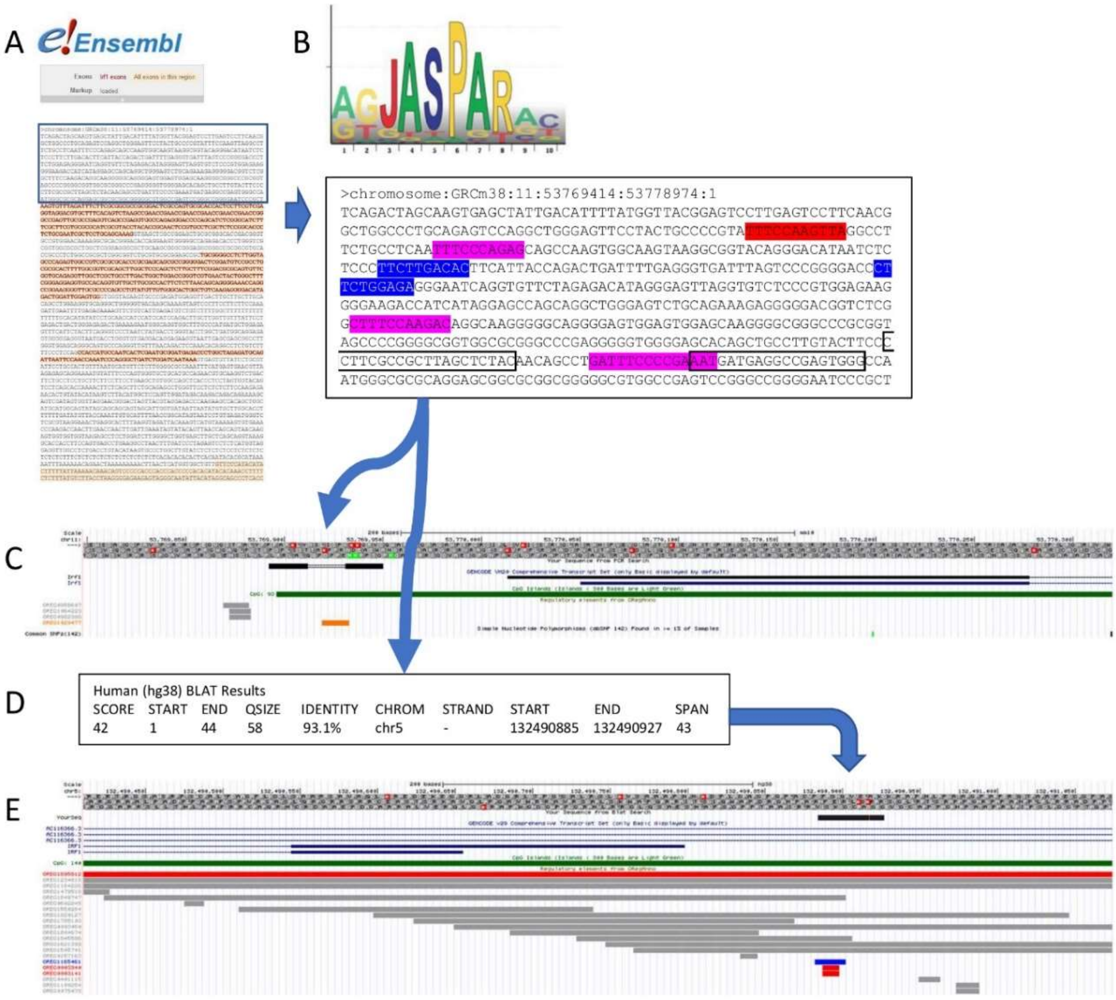
2.4.2. Oligonucleotide primer sequences.

Table 2.4 lists the sequences of the primers used for ChIP-qPCR. Other primers used (*Bcl3*, *Bcl6*, *Icam1*, *Il4r* super enhancer, *JunB*, *Stat3*, *Zeb1*) were designed using the "Custom Plus Taqman Assay Design Tool" from ThermoFisher. Promoter sequences (P) (see Figure 2.5) are sequences with known binding sites, whilst downstream sequences (DS) act as negative controls with no known binding sites.

Table 2.4: Oligonucleotide primer sequences used for chromatin immunoprecipitation.

Promoter sequences (P) with known STAT binding sites, whilst downstream sequences (DS) are utilised as negative controls. TaqMan probes utilise a third primer that binds in the middle of the sequence that is incorporated into the new synthesised copy releasing the fluorescent marker. This third primer increases the specificity of reported qPCR amplification

Target	Forward	Probe	Reverse
Irf1 P	CCTTCGCCGCTTAGCTCTAC	ACAGCCTGATTTCC	CCCACTCGGCCTCATCATT
Irf1 DS	GCCTTGGCGTGACTCTTGAC	ATCTATTAGAAACGCCACCTAA	ACATGACCAAACACCATTTAGCA
Socs3 P	CTCCGCGCACAGCCTTT	TGCAGAGTAGTGACTAAA	CCGGCCGGTCTTCTTGT
Socs3 DS	GGGTAATTCCTGCCGTCTGA	TCTGACCAGAATATGC	CATTTCCTTCGCAAACTTGCT



```
>chromosome:GRCm38:11:53769414:53778974:1
TCAGACTAGCAAGTGAGCTATTGACATTTTATGGTTACGGAGTCCTTGAGTCCTTCAACG
GCTGGCCCTGCAGAGTCCAGGCTGGGAGTTCCTACTGCCCCGTATTTCCAAGTTAGGCCT
TCTGCCTCAATTTCCCAGAGCAGCCAAGTGGCAAGTAAGGCGGTACAGGGACATAATCTC
TCCCTTCTTGACACTTCATTACCAGACTGATTTTGAGGGTGATTTAGTCCCGGGGACCCT
TCTGGAGAGGGAATCAGGTGTTCTAGAGACATAGGGAGTTAGGTGTCTCCCGTGGAGAAG
GGGAAGACCATCATAGGAGCCAGCAGGCTGGGAGTCTGCAGAAAGAGGGGGACGGTCTCG
GCTTTCCAAGACAGGCAAGGGGGCAGGGGAGTGGAGTGGAGCAAGGGGCGGGCCCGCGGT
AGCCCCGGGGCGGTGGCGCGGGCCCGAGGGGGTGGGAGCACAGCTGCCTTGTACTTCCC
CTTCGCCGCTTAGCTCTACAACAGCCTGATTTCCCCGAAATGATGAGGCCGAGTGGGCCA
ATGGGCGCGCAGGAGCGGCGCGGCGGGGGCGTGGCCGAGTCCGGGCCGGGGAATCCCGCT
```

Human (hg38) BLAT Results

SCORE	START	END	QSIZE	IDENTITY	CHROM	STRAND	START	END	SPAN
42	1	44	58	93.1%	chr5	-	132490885	132490927	43

Figure 2.5: Rational behind primer design for chromatin immunoprecipitation.

In this example for Irf1. ***A;*** *The promotor sequence is extracted from Ensembl.* ***B;*** *the promoter sequence is scanned for STAT motifs in Jaspar, STAT1 shown in red, STAT3 in blue, and both STAT1 and STAT3 in purple.* ***C;*** *in-silico PCR identifies only the 1 amplified sequence, which overlaps the Irf1 promoter. As can be seen, the regulatory elements of the mouse genome are relatively sparse, with a STAT4 binding site (orange) identified within the primer amplification sequence.* ***D;*** *Blasting this sequence against the Human genome identifies one match in the promoter of IRF1.* ***E;*** *the Human genome is much more complete for regulatory elements, as can be seen the blasted sequence overlaps STAT1 (red) and STAT3 (blue) binding sites*

2.5. ChIP-seq.

DNA was quantified using Qubit (ThermoFisher Scientific), and enrichment checked using qPCR as outlined previously.

Library preparation follows the protocol outlined in the Illumina TruSeq ChIP sample preparation guide (Illumina 15023092 Rev. B) with a minor adaptation – performing the size selection step after PCR amplification of the fragments.

In brief, the TruSeq ChIP library preparation has 6 stages; end repair, A-tail, ligate adapters, library amplification, size selection, and library validation, illustrated in Figure 2.6.

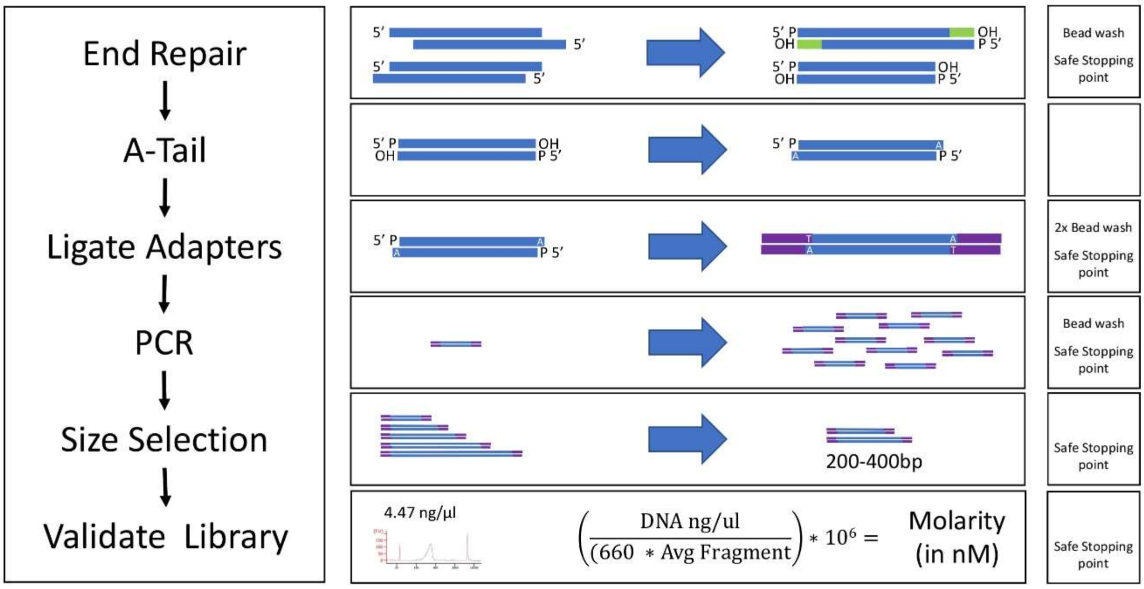


Figure 2.6: Workflow for library preparation of ChIP samples.

End repair produces blunt end fragments by removing 3' overhangs using 3' to 5' exonuclease and fills in 5' overhands via polymerase. A-tailing adds adenine overhangs to the 3' end that complements the 3' thymine overhang on the adapters. Adapters contain index sequences that allow samples to be multiplexed, they also contain sequences that facilitate binding to the flow cell of the sequencer. Libraries are amplified with 18 cycles of PCR, before restriction fragment sizes to 200-400 base pairs on the BluePippin. Libraries are quantified using Qubit and DNA high sensitivity bioanalyzer chip. Bead washes purify DNA products utilising AMPure XP beads that bind DNA fragments larger than 100 base pairs. Safe stopping points allow for the protocol to be broken up over multiple days with samples being stored at -20°C.

End repair: DNA fragmentation can lead to overhanging sequences; therefore, samples are treated with a 3' to 5' exonuclease to remove 3' overhangs and leave the sample with blunt ends for the next step.

A-tail: Samples have the 3' end adenylated to provide a complementary overhang for the adapters to ligated to.

Ligate adapters: Samples have an adapter sequence ligated to them that identifies the samples with a barcode sequences, as well as complementary sequences needed for hybridisation to the flow-cell when sequencing.

Library amplification: PCR amplification of the samples allows for enrichment of DNA fragments containing the adapter molecule ligated to both ends of the sequence. See Table 2.5 for PCR parameters.

Table 2.5: PCR conditions for library amplification as specified in the Illumina TruSeq ChIP library preparation guide.

	Stage	Temperature	Time (seconds)
	Hold	98°C	30
PCR (18 cycles)	Dentature	98°C	10
	Anneal	98°C	30
	Extend	72°C	30

Size selection: Restriction of DNA fragment length was performed using BluePippin (Sage Science) (2% agarose cassette, 200-400bp), rather than agarose gel excision followed by purification listed in the Illumina protocol.

Library validation: Library quality was assessed on High Sensitivity DNA Bioanalyzer (Agilent) and quantified using Qubit.

Libraries were standardised to 10 nM, and pooled before next generation sequencing on an Illumina HiSeq 4000. Bioinformatic analysis of the data is covered in Chapter 3.

2.6. Public Datasets.

Public databases (NCBI GEO, EBI Array Express, Immport) were trawled for datasets that containing arthritic synovial samples. Table 2.6 outlines all the identified microarray datasets matching this criterion, whilst Table 2.7 represents RNA-seq datasets. The development of bioinformatic analysis pipelines associated with these datasets is discussed in Chapter 3.

Accession	Platform	n	Description
GSE48780	Affymetrix HGU133plus2	83	2 Cohorts of RA synovium from joint resection (Cohorts of 49 and 34 patients)
GSE45867	Affymetrix HGU133plus2	40	Synovial biopsies before and after treatment with either Tocilizumab (12 patients: 24 samples) or Methotrexate (8 patients: 16 samples)
GSE24742	Affymetrix HGU133plus2	24	Synovial biopsies before and after treatment with Rituximab (12 patients)
GSE36700	Affymetrix HGU133plus2	25	Synovial biopsies from different forms of arthritis, 5 Osteoarthritis, 7 Rheumatoid Arthritis, 4 Systemic Lupus Erythematosus, 5 Microcrystalline arthritis, 4 Seronegative Arthritis.
GSE77298	Affymetrix HGU133plus2	23	Synovial biopsies from end-stage rheumatoid arthritis (16 samples) and healthy controls (7 samples)
GSE15602	Affymetrix HGU133plus2	11	Synovial biopsies after treatment with Adalimumab
GSE38064	Affymetrix HGU133plus2	12	Synovial biopsies from patients which were CD21L+IL-17A+ (7 samples) or CD21L-IL-17A- (5 samples)
GSE55457	Affymetrix HGU133A	33	Jena - Synovial biopsies from Healthy (10 samples), Rheumatoid Arthritis (13 samples), Osteoarthritis (10 samples)
GSE55584	Affymetrix HGU133A	16	Leipzig - Synovial biopsies from Rheumatoid Arthritis (10 samples), Osteoarthritis (6 samples)
GSE55235	Affymetrix HGU133A	30	Berlin - Synovial biopsies from Healthy (10 samples), Rheumatoid Arthritis (10 samples), Osteoarthritis (10 samples)
GSE12021	Affymetrix HGU133A/B	57	Synovial biopsies from Healthy (13 samples: 9A, 4B), Rheumatoid Arthritis (24 samples: 12A, 12B), Osteoarthritis (20 samples: 10A, 10B
GSE1919	Affymetrix HGU95A	15	Synovial biopsies from Healthy (5 samples), Rheumatoid Arthritis (5 samples), Osteoarthritis (5 samples)
GSE2053	HUMAN UNIGENE SetI Part 1	8	Synovial biopsies from Healthy (4 samples), Rheumatoid Arthritis (4 samples)
GSE39340	Illumina HumanHT-12 V4.0 expression beadchip	22	Synovial biopsies from Ankylosing Spondylitis (5 samples), Rheumatoid Arthritis (10 samples), Osteoarthritis (7 samples)
E-TABM-104	KTH H. sapiens 29.8k cDNA v2/ KTH H. sapiens 30.5k cDNA array v1	32	Synovial biopsies before and after treatment with infliximab (10 patients)
GSE21537	KTH H. sapiens 30.5k cDNA array v1	62	RA synovial biopsies before infliximab treatment
GSE13026	INSERM Homo sapiens 14K array Liverpool3	45	Synovial biopsies from Healthy (21 samples), early Rheumatoid Arthritis (12 samples), late Rheumatoid Arthritis (12 samples)
GSE3698	Human Unigene3.1 cDNA Array 37.5K v1.0	48	Synovial biopsies from Osteoarthritis (19 samples), Rheumatoid Arthritis (18 samples), Pigmented Villonodular Synovitis (11 samples)
GSE3848	LC-14	31	Synovial biopsies from Osteoarthritis (9 samples), Rheumatoid Arthritis (22 samples)

Table 2.6: Microarray datasets identified by searching online databases for samples that contained the terms “arthritis” “synovial” “synovium”.

Accession	Platform	# samples	Description
SDY998	Illumina HiSeq 2500	22	AMP Rheumatoid Arthritis Arthroplasty Phase 1
SDY999	Illumina HiSeq 2500	34	AMP Rheumatoid Arthritis Synovial Phase 1
SDY1299	Illumina HiSeq 2500	45	Identification of Three Rheumatoid Arthritis Disease Subtypes by Machine Learning Integration of Synovial Histologic Features and RNA Sequencing Data
E-MTAB-6141	Illumina HiSeq 2500	154	Pathobiology of Early Arthritis Cohort (PEAC), synovial biopsies (87 samples) and blood (62 samples) samples from treatment naïve rheumatoid arthritis patients
GSE89408	Illumina HiSeq 2000	218	Synovial biopsies from Healthy (28 samples), Rheumatoid Arthritis (152 samples), Osteoarthritis (22 samples), Arthralgia (10 samples), Undifferentiated arthritis (6 samples)
GSE97165	Illumina HiSeq 2000	38	Synovial biopsies before and after treatment with triple DMARDs (19 patients)

Table 2.7: RNA-seq datasets identified by searching online databases for samples that contained the terms "arthritis" "synovial" "synovium".

3. Development of analytical bioinformatic methods.

This thesis has made extensive use of bioinformatic analyses to identify the gene signatures that discriminate the forms of synovitis in rheumatoid arthritis patients. This chapter outlines the principles of the approaches utilised throughout this thesis, with specific adaptations described in their respective chapters.

3.1. Microarray analysis.

Figure 3.1 summarizes the workflow utilised across the various steps associated with microarray analysis, the details of which are explored in more detail below.

3.1.1. Microarray platforms.

As listed in Table 2.4, there have been multiple microarray platforms used to identify the transcriptome of synovial samples from rheumatoid arthritis patients. Ultimately, these can be broken down to two categories, single and dual-channel detection methods. Single channel microarrays load a single sample per microarray chip, whist dual-channel methods utilise paired samples (e.g., healthy vs. diseased tissues) that are individually coded with fluorescent labels and provide information based on the relative abundance of transcripts.

In this thesis, we have focussed on utilising the more common HGU133A/B and HGU133plus2 platforms for their well-annotated probe-sets and relatively consistent behaviour across samples.

3.1.2. Sample metadata preparation.

Metadata was obtained from information deposited in open access repositories and supporting data presented in the original study publications. Data was compiled as a table and saved as a character separated value (CSV) file that listed file name, sample ID, and whatever additional information was available (e.g., Disease Activity Score, response to therapeutics, histological data).

3.1.3. Normalisation.

Samples were normalised using the Robust Multi-Array (RMA) average expression measure from the affy package to allow comparison of the samples across the experiment(118). RMA relies on three steps termed background correction, quantile normalisation, and median polishing (summarised in Figure 3.2).

- Background correction utilises the probe-match mismatch to correct each microarray individually for background noise from aberrant binding.

- Quantile normalisation adjusts the distributions of all samples to normalise these across all samples being explored. This is performed ranking all the probes from largest to smallest, taking the average for each rank and returning this value for each sample.
- Median polish then summarises the probes and returns a single intensity value for each probe-set. For each probe-set there are multiple probes, these are polished by subtracting the row medians (per probe) and then column medians (per sample), and this is repeated until the medians converge with a maximum of 5 iterations.

3.1.3.1. Batch effects and correction attempts.

Many of the datasets used in this study lack 'appropriate' baseline controls and are entirely comprised of transcriptomic data from diseased samples. Where data was available from healthy controls, these were often insufficiently powered to generate meaningful statistical determinants. These issues represented a major challenge for this study. Therefore, attempts were made to combine microarray datasets from various patient cohorts containing data from healthy and diseased tissues (see Chapter 4.5.1).

As a good example of these batch effects, 3 linked datasets (Berlin – GSE55235, Leipzig – GSE55584, Jena – GSE55457) representing a total of 79 patient samples from rheumatoid arthritis, osteoarthritis, and healthy controls were combined. As demonstrated in Figure 3.3A, unsupervised clustering revealed no association with the individual disease states. Instead, samples clustering mapped almost entirely with the site where the experiment was conducted (Figure 3.3B). In an attempt to correct for these batch effects, three form correct analysis were applied to the datasets– removeBatchEffect from the limma package(9), frozenRMA(119), and Combatting Batch effects (ComBat)(120). These methods initially appeared to fix the batch issue. For example, ComBat correction of the Berlin-Leipzig-Jena datasets in Figure 3.3C. However, further investigation found that these correction methods restricted the downstream analysis of biological behaviours within individual datasets.

3.1.4. Stratification of disease pathologies.

Identification of patient pathology is the principle aim of this thesis, to achieve this several methods were utilised to group samples for downstream analysis. Initial stratification utilised unsupervised clustering to group patients into the 3 pathologies of rheumatoid arthritis and is described extensively in Chapter 4. Following experiments made use of metadata or classifier results to stratify these groups for further analysis

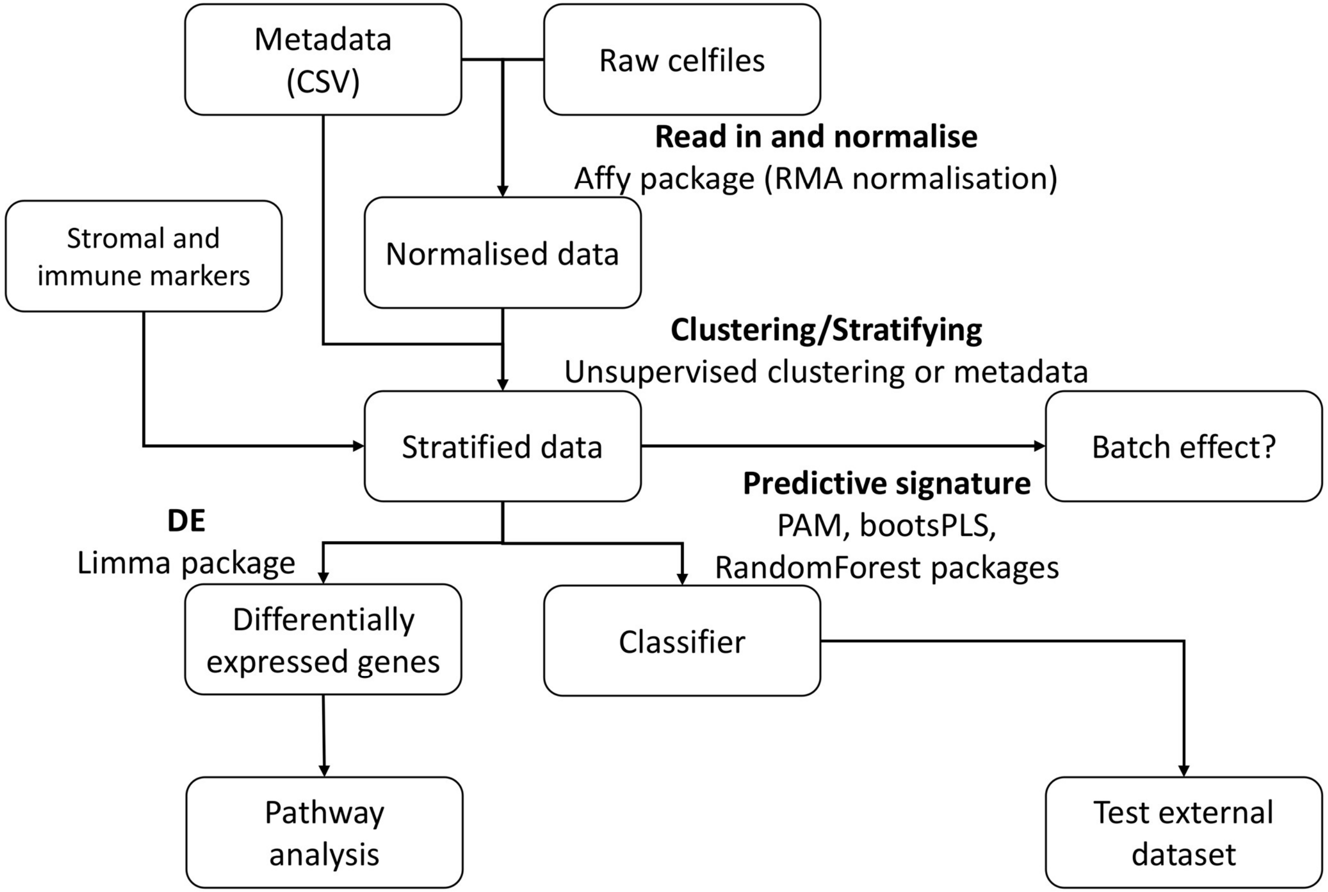


Figure 3.1: The workflow utilised in microarray analysis.

This basic workflow has been utilised to identify pathways affected in disease and create classifiers capable of discriminating disease pathologies.

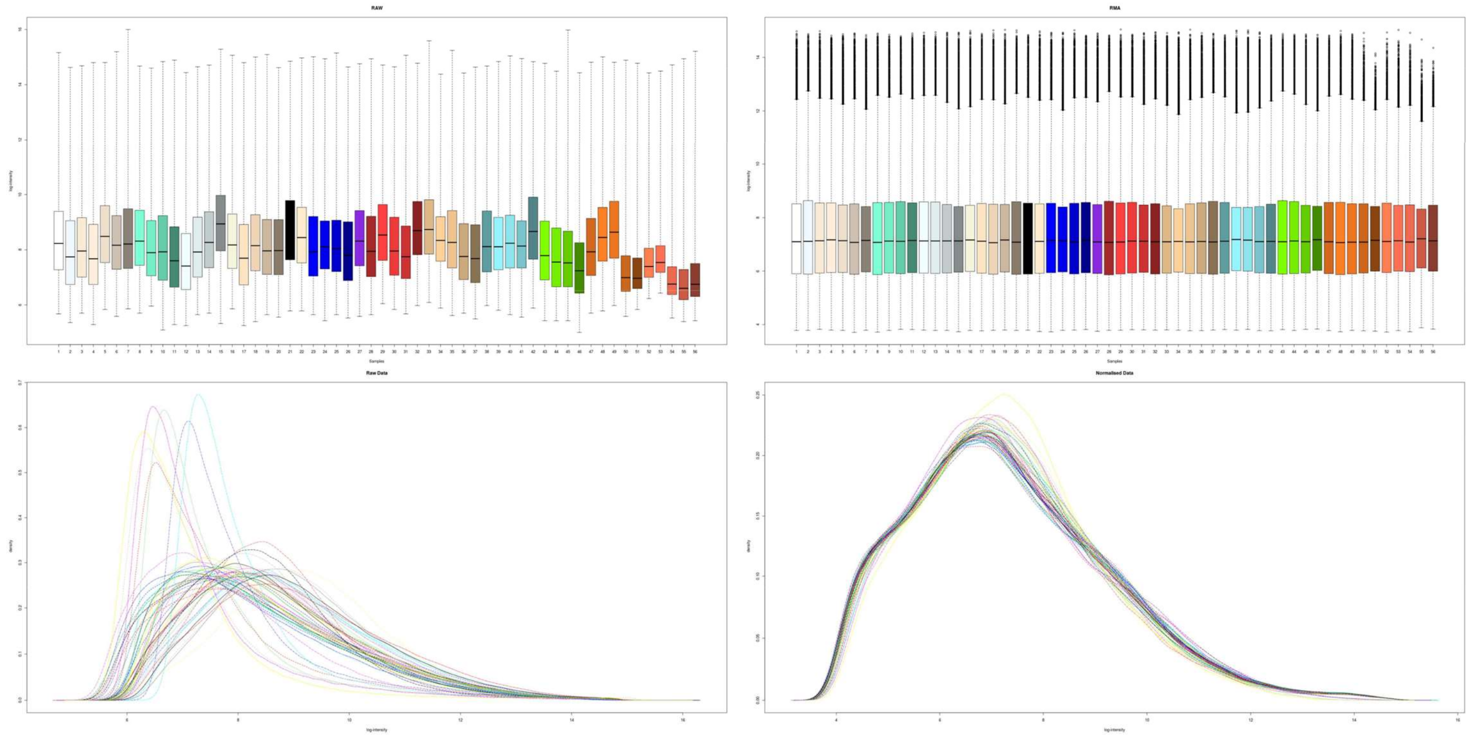


Figure 3.2: Boxplot and density curves of expression values before and after RMA normalisation.

The data plotted here represents cohort1 and the healthy controls discussed in Chapter 4. As can be seen in the raw values the 7 healthy controls at the end of the boxplots are fundamentally different to cohort1, and this is even more clear in the density plots. After RMA normalisation this distribution becomes comparable between the 2 datasets.

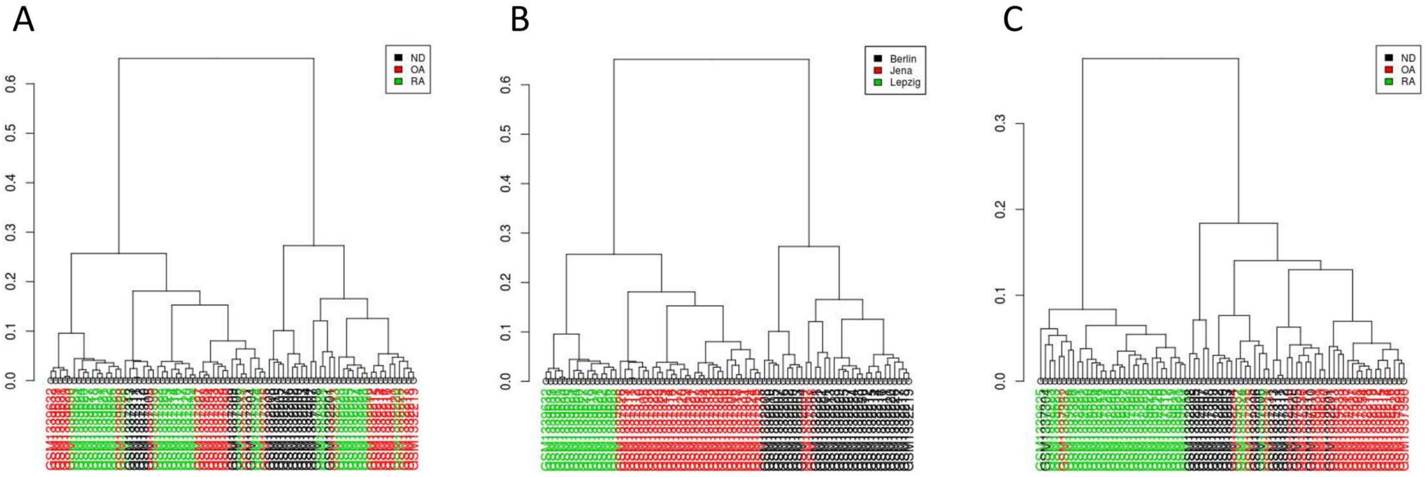


Figure 3.3: Illustrating the confounding effects associated with batches.

Three datasets or synovial samples derived from patients with rheumatoid arthritis, osteoarthritis, and no disease, from clinics in Berlin, Leipzig, and Jena. ***A****) Unsupervised clustering of the datasets, and colouring by disease revealed no association between samples.* ***B****) Changing the colouring to reflect which research centre performed the experiments shows that this factor dominates the differences between samples.* ***C****) Correcting for batch effects using ComBat results in disease being the primary differentiating factor.*

3.1.5. Immune and stromal cell marker lists

Initial attempts to quantify the cellular composition of synovial tissue samples based on the transcriptional profile of a biopsy used the Cellmix package(121). However, this method was not well suited for cellular deconvolution in a complex tissue biopsy. Instead, gene marker lists(122–125) derived in this package were combined to provide a more comprehensive cell-type specific set of markers. These lists were further refined to consolidate any duplicate transcripts. To support the analysis of stromal cells, a bespoke fibroblast-related gene list was generated using transcriptomic data from differentiated synovial fibroblasts from rheumatoid arthritis and osteoarthritis patients, bone marrow fibroblasts and skin fibroblasts(126).

The utility of these marker sets is shown Figures 3.4 and 3.5, which demonstrates their performance against 4 independent datasets of purified immune cells from PBMC.

3.1.6. Generating a predictive gene signature.

Details describing the generation of specific gene signatures is outlined in Chapter 4 and 6. However, several adaptation were made to the sparse Partial Least Squares (sPLS) approach, which was used to identify the optimal number of genes utilised in each component of the classifier.

An initial exploratory attempt utilised an early version of the bootsPLS package(14) which necessitated running in single-threaded manner. As explained in Section 4.5.6, this initial attempt that identified a 13 gene signature, took over 18 hours to run with only 15 iterations. For the main analysis to be robust it needed many more iterations, therefore we investigated methods for improving computation time. With an updated version of the package, this improved compute time and when combined with multi-threading allowed many more iterations to be run using the HAWK supercomputer. To optimise the usage of compute resources, we identified a bottleneck that showed a maximal thread utilisation of the package, where going beyond 6 threads did not reduce compute time per iteration (Figure 3.6). Therefore, multiple runs were performed in parallel utilising 6 threads for maximal performance. As there is some variability in the number of components selected in the final model, a random seed was assigned to ensure that the data was reproduceable. All models generated are then denoted as fitX, where X represents the random seed utilised.

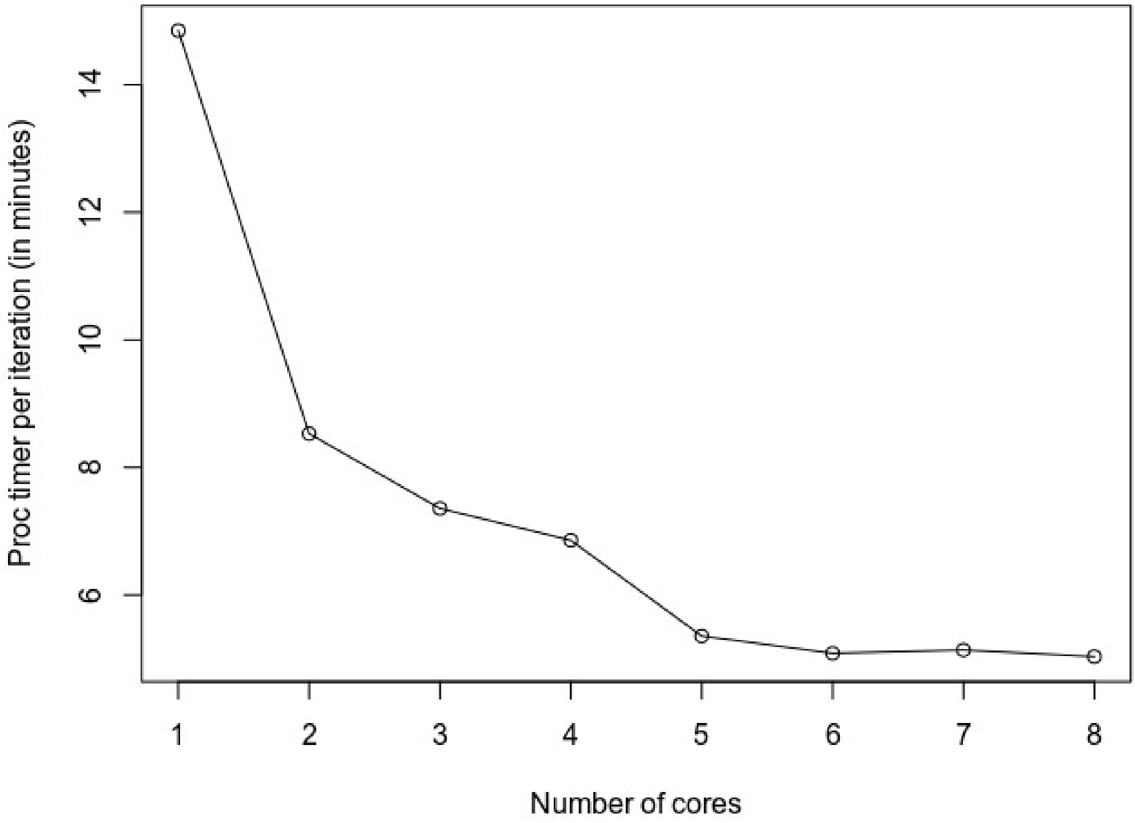


Figure 3.4: Compute time and optimisation of signature generation.

This illustrates the time (in minutes) taken to complete one iteration of bootsPLS on the training dataset (GSE48780). Compute time flattens out after 6 threads, therefore computation of the final signatures was performed by running multiple scripts in parallel on the HAWK supercomputer, each using 6 threads for maximal efficiency.

3.1.1. Differential gene expression.

All experiments utilised limma to identify differentially expressed genes using an empirical Bayes method(9). The details of each comparison are explained in their respective chapters, and unless stated otherwise used Bonferroni correction to adjust p-values for further analysis.

3.1. RNA-seq analysis.

This thesis utilised publicly available RNA-seq datasets, and was not focussed on differential gene expression, and therefore has a more limited pipeline, focussed on mapping and quantifying the reads for stratification. Certain datasets were obtained from repository sites as pre-processed tables and therefore didn’t utilise this analysis.

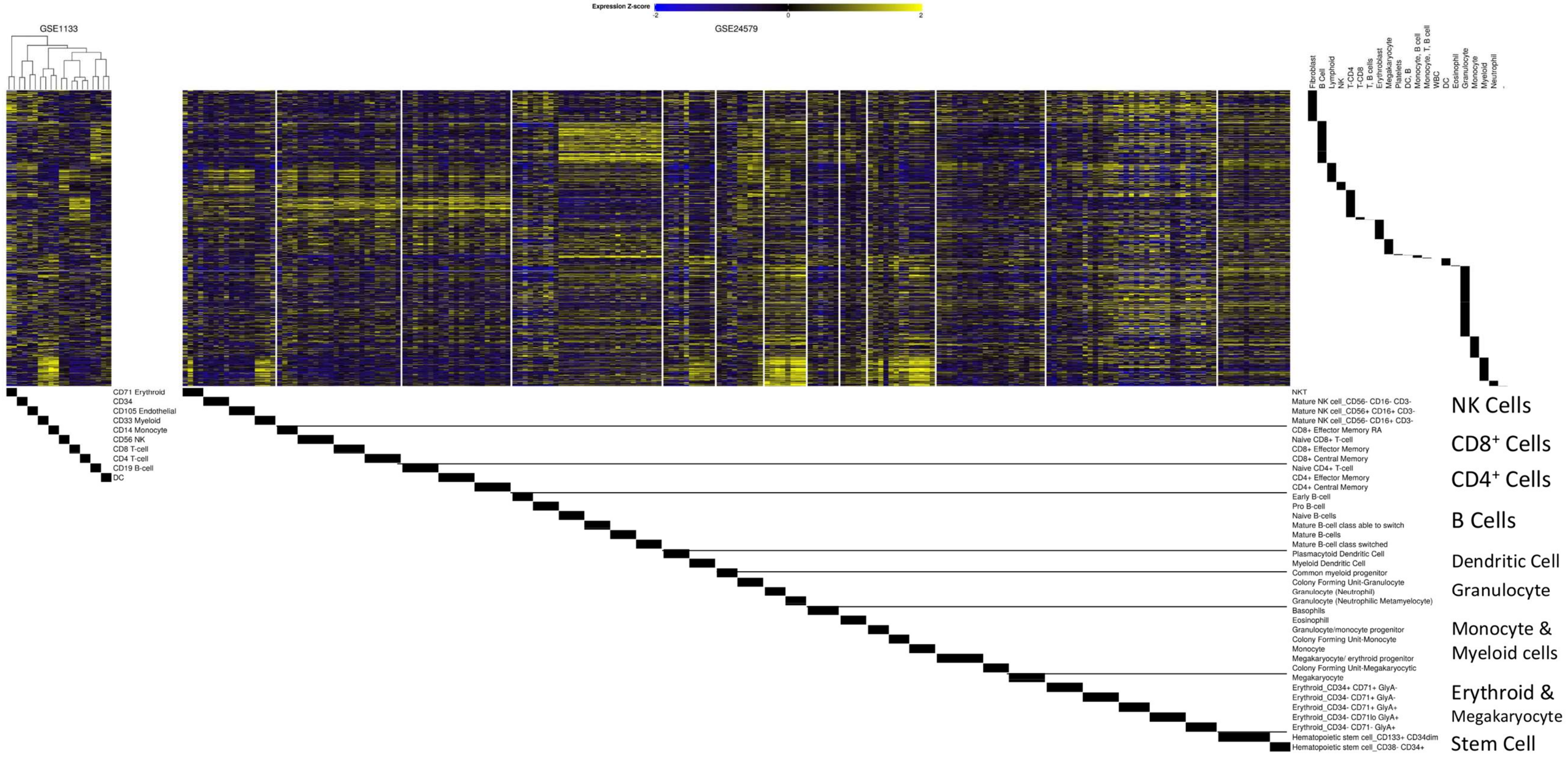


Figure 3.5: Immune and stromal markers performance on purified cells (HGU133A platform).

Two purified immune cell datasets (GSE1133 & GSE24579) were restricted to immune and stromal markers as outlined in Chapter 3.1.5. These markers robustly identify lymphoid cells, and preform well with myeloid cells.

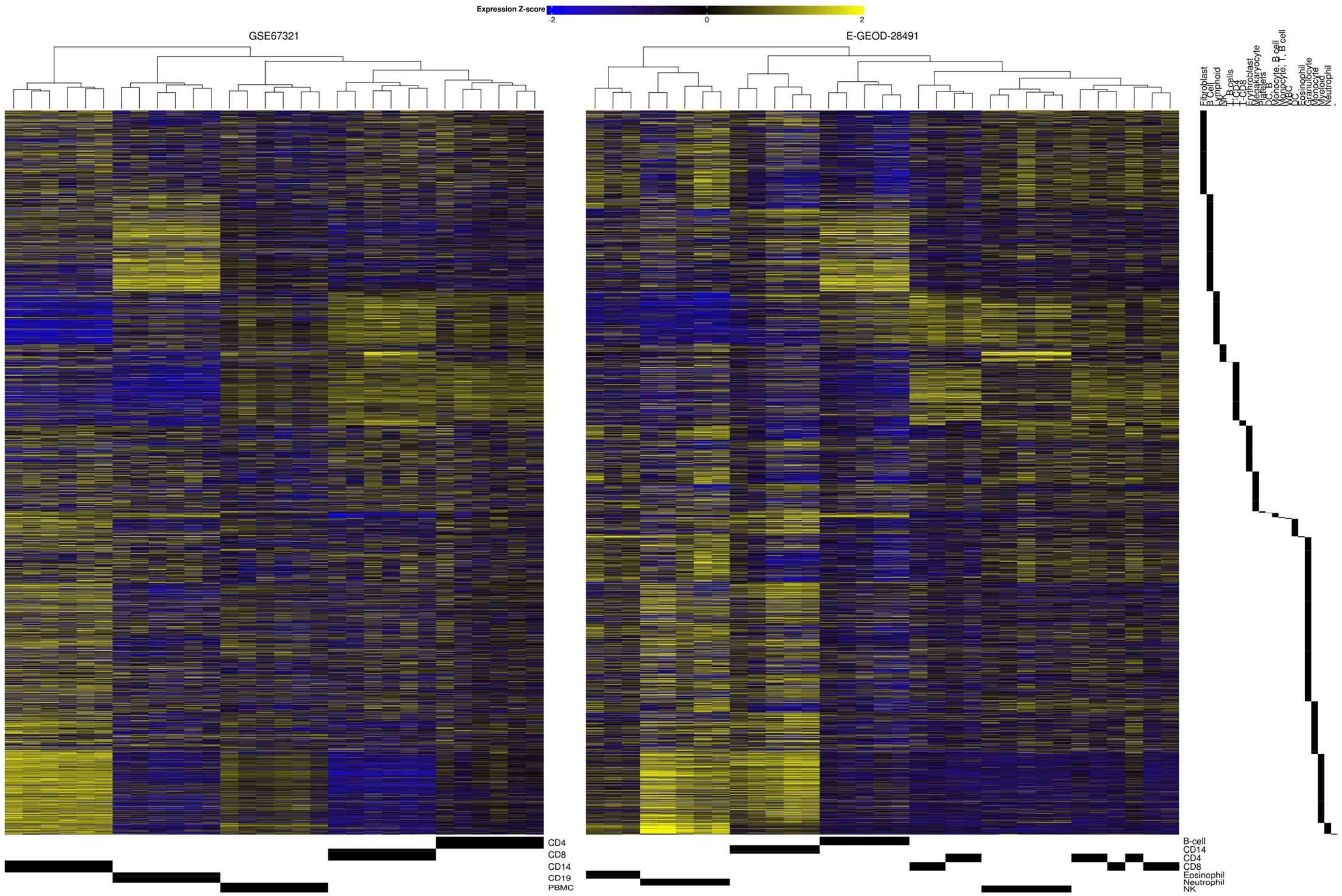


Figure 3.6: Immune and stromal markers performance on purified cells (HGU133plus2 platform).

Two purified immune cell datasets (GSE67321 and E-GEOD-28491) were restricted to immune and stromal markers as outlined in Chapter 3.1.5. These markers robustly identify lymphoid and myeloid cell populations

3.1.1. Mapping reads to the genome.

One issue that was raised when testing the gene signatures derived in Chapter 4, was the identification of signature genes that were found in human alternative sequences. Figure 3.7 illustrates how these alternative sequences, from allelic sequences to fix patches are associated with the chromosome, with examples of the contigs that make up the alternative sequences. In the case of one of these genes - LILRA3, this was found on chromosome 19, in a region with 4 human alternative sequences. This results in this gene having 4 separate ENSEMBL identifiers (ENSG00000273884, ENSG00000275841, ENSG00000276175, ENSG00000278046) associated with the 4 different alternative sequences that cover this part of the genome. To address this, the RNA datasets were mapped to the Gencodes GRCh38.p12 genome using Burrows-Wheeler Aligner (BWA)(127). Gencodes annotation file amalgamates the RefSeq, Ensembl and alternative sequences, providing comprehensive coverage.

3.1.2. Generating a preditive gene signature in early arthritis.

This utilised the same basic methodology as outlined in section 3.1.6, but with some adaptations required for utilisation with RNA-seq data.

For use with the PEAC dataset, the data needed to be transformed from FPKM to TPM, as the low expression levels interfered with the analysis. Moreover, the data needed extensive filtering to obtain a complete analysis, requiring the removal of near-zero variants using the packages build-in function, further filtering was performed to remove all pseudoautosomal region and any genes that had a 0 in more than 25% of samples

3.2. ChIP-seq analysis.

Figure 3.8 summarises the ChIP-seq workflow used to identify peaks in the data.

3.2.1. Quality control of reads.

After sequencing, reads were demultiplexed and trimmed to remove adapter sequences. The quality of these reads were then assessed using fastqc. Duplicate reads were flagged prior to mapping.

3.2.2. Mapping reads to the genome.

Reads were mapped to GRCm38.84 (mm10) using BWA for both marked and removed duplicate samples.

3.2.3. Peak calling.

Duplicates were removed and peaks were called using Model-based Analysis of ChIP-seq (MACS2)(128) at three q-value thresholds (0.1, 0.05, 0.01). Bed files generated by peak calling were then analysed to identify overlaps with genes using the Peak Annotation and

VISualisation (PAVIS)(129) service using the default parameters of 5kb upstream and 1kb downstream of the transcription start site.

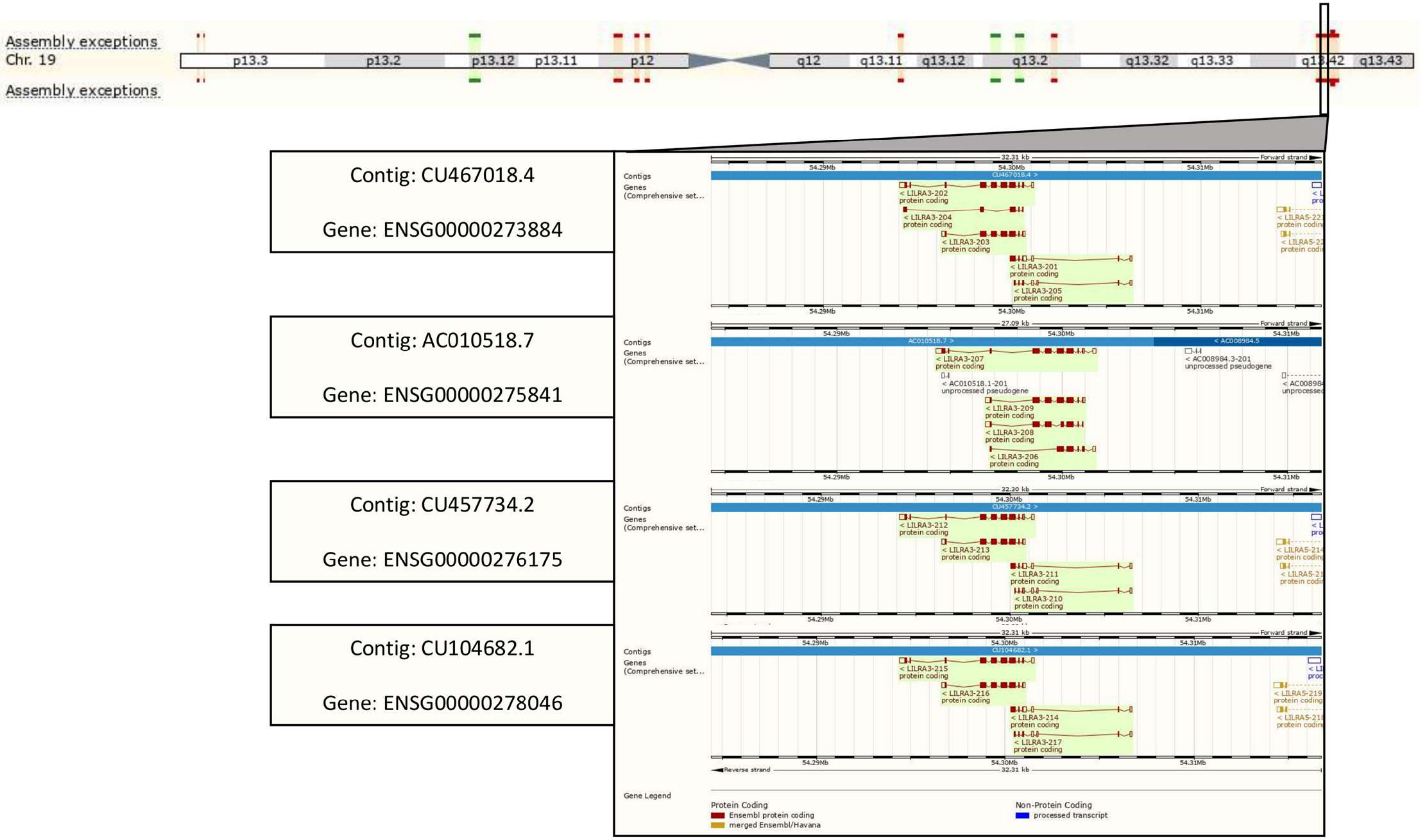

Assembly exceptions
Chr. 19
Assembly exceptions
p13.3
p13.2
p13.12
p13.11
p12
q12
q13.11
q13.12
q13.2
q13.32
q13.33
q13.42
q13.43
Contig: CU467018.4
Gene: ENSG00000273884
Contig: AC010518.7
Gene: ENSG00000275841
Contig: CU457734.2
Gene: ENSG00000276175
Contig: CU104682.1
Gene: ENSG00000278046
Contigs
Genes (Comprehensive set...
< LILRA3-202 protein coding
< LILRA3-204 protein coding
< LILRA3-203 protein coding
< LILRA3-201 protein coding
< LILRA3-205 protein coding
< LILRA3-207 protein coding
< AC010518.1-201 unprocessed pseudogene
< LILRA3-209 protein coding
< LILRA3-208 protein coding
< LILRA3-206 protein coding
< AC008984.3-201 unprocessed pseudogene
< LILRA3-212 protein coding
< LILRA3-213 protein coding
< LILRA3-211 protein coding
< LILRA3-210 protein coding
< LILRA3-215 protein coding
< LILRA3-216 protein coding
< LILRA3-214 protein coding
< LILRA3-217 protein coding
54.29Mb
54.30Mb
54.31Mb
32.31 kb
27.09 kb
32.30 kb
Forward strand
Reverse strand
Gene Legend
Protein Coding
Ensembl protein coding
merged Ensembl/Havana
Non-Protein Coding
processed transcript

Figure 3.7: Human alternative sequences associated with chromosome 19.

Alternative allelic sequences associated with haplotypes are depicted in red, fix patches to the reference genome in green. This figure highlights the 4 Human alternative sequences associated with LILRA3, resulting in 4 different ENSEMBL gene ID's expanded out from the main chromosome.

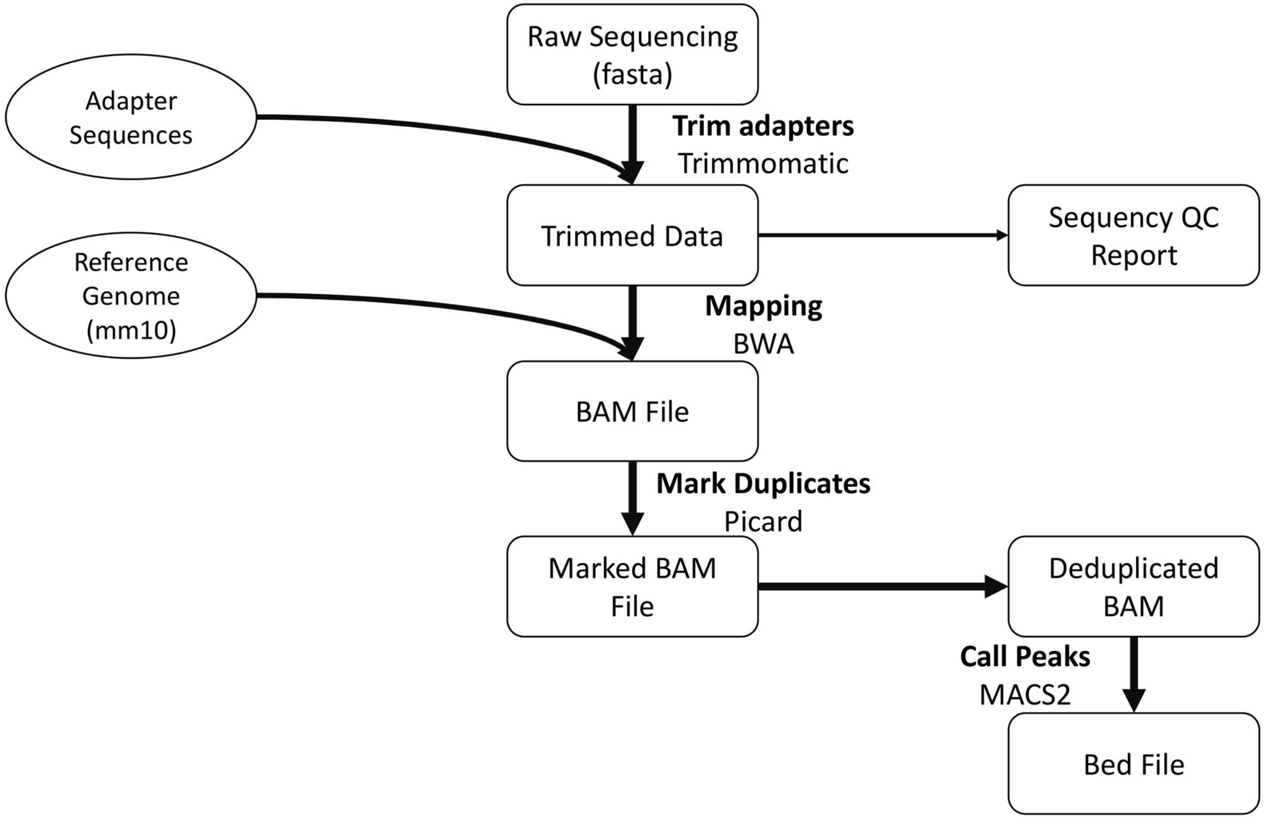


Figure 3.8: Workflow for the ChIP-seq pipeline.

Raw sequencing data is trimmed to remove multiplexing adapters before being mapped to the reference genome. Peaks are identified using the non-duplicated reads, and can be passed downstream to for further analysis to identify associated gene

4. Stratification of synovial pathology according to transcriptional gene expression.

4.1. Introduction.

Clinical experience shows that early diagnosis and treatment of rheumatoid arthritis prevents irreversible joint damage and offers the best opportunity for drug-free remission. However, ~40% of patients fail to show adequate response to standard biological drug therapies(16,52). This lack of efficacy potentially reflects the clinical heterogeneity of disease seen in patients with RA. For example, patients with RA often show considerable variability in the rate of disease progression and severity. These differences in the clinical presentation of RA is epitomised by studies of synovial joint inflammation (synovitis), which shows that the histological features of disease vary from patient-to-patient. Whilst routine blood tests – e.g. measurements of the acute-phase reactants like CRP and ESR, as well as autoantibodies such as Rheumatoid Factor and anti-citrullinated protein antibody (ACPA) – are important tools in the clinical diagnosis of RA, they provide limited information on the type of synovial pathotype seen in patients(71,75,130,131). Thus, there is a need to understand the inflammatory processes driving these distinct forms of pathology.

Investigations into the molecular basis of disease heterogeneity in RA have significantly benefitted from advances in ultrasound-guided biopsy sampling techniques, which have allowed the isolation of synovial tissue biopsies at an early stage of disease progression(65,69,132,133). Histological examination of these biopsies has identified characteristic features of RA synovitis that include hypertrophy of the synovial lining layer, neo-angiogenesis, and the infiltration of leukocytes associated with the control of innate and adaptive immune responses(134,135). However, the pattern of synovitis identified in RA patients varies considerably, and synovitis is broadly classified according to defined histological features(70,136). These are termed *Follicular* (often termed lymphoid-rich synovitis), *Diffuse* (often termed myeloid-rich synovitis), and *Pauci immune* (often termed fibroblast-rich synovitis), see Section 1.3.5 for more in-depth exploration of the characteristics of the different pathologies. The definition of these distinct forms of synovitis suggests that the underlining inflammatory mechanisms of joint disease differ between RA patients and is likely to influence both the rate and severity of disease onset as well as the response to biological drug intervention(137,138). Any improvement in the diagnosis or stratification of these different pathologies will ultimately improve clinical decisions on the best course of therapy for a distinct patient group(139).

Advances in high-throughput transcriptomic technologies have allowed investigators to explore gene expression associated with the development of synovitis(140,141). Previous work has predominantly focussed on identifying the characteristics that discriminate RA from similar diseases such as osteoarthritis or looking at the response to therapeutics(142,143). In the majority of these cases, the studies do not take into account pathology and how this may have a confounding effect on the experiment(144,145). A small number of studies have utilised transcriptomic data in conjunction with the defined histological features to identify unique molecular phenotypes specific to each of the pathologies(73,130).

Just as the cause of RA is unknown, so too are the factors that lead to the differentiation of disease and therefore clinical heterogeneity(32). Consequently, investigation of the pathways associated with the different pathologies may identify commonalities that underly them and indicate the mechanisms controlling and potentially initiating the disease(66).

4.2. Hypothesis.

Previous investigations have demonstrated that the pathologies have their own unique molecular phenotype, yet these analyses have been performed on highly stratified patient samples(73,130). However, in routine clinical practice, many patients will display discrete patterns of synovitis that do not adhere to one of the three pathologies and therefore represent a “grey area” of patients.

By identifying a characteristic profile that defines the three different pathologies, it is possible to characterise the profile of these “grey area” patients and determine the similarities with the distinct pathologies. Furthermore, a defined archetypical profile will allow the retrospective analysis of datasets which, in general, have very limited metadata respecting the patients. Retrospective stratification of these datasets will open the possibility of increased power to detect the effects of therapeutics, where previously the increased variation caused by not defining the pathologies reduced the ability to see any effects.

4.3. Aims.

This chapter will explore the biology underlying the three pathologies, looking at similarities and differences between them. It will also demonstrate computational methods for identifying the different pathologies and validating them using immune marker profiles.

4.4. Materials and Methods.

4.4.1. Reading in data, metadata.

Raw sample data was obtained from the NCBI Gene Expression Omnibus (GEO) repository under the accession GSE48780, this is the data associated with the Dennis *et al.* (2014) paper, which identified 4 major phenotypes with distinct gene expression signatures.

GSE48780 contains 83 samples from 2 cohorts of RA patient-derived synovial tissue from arthroplasty and/or synovectomy. Metadata provided in the series matrix file contains the following additional information for cohort 1 (labelled batch) and is otherwise missing for the second cohort: gender, joint location, inflammation, and batch. For confirmation of this, the scan date of the individual .cel files was extracted, as can be seen in Table 4.1 the 2 cohorts were processed in 2008 and 2010 respectively.

Samples from the first cohort were read in and normalised as described in Section 3.1.3.

Table 4.1: Metadata for GSE48780 extracted from series matrix file and scan date from the celfile.*

The first cohort of 49 patients utilised in the Dennis et al. paper is clearly distinguished by the level of detail in the metadata. This is also validated by the difference in scanning date of the microarray, Cohort 1 were all scanned in December 2008, whilst Cohort 2 was in November 2010. Samples belonging to the second cohort are highlighted in blue.

Accession #	*Gender*	*Joint Location*	*Inflammation*	*Batch*	*ScanDate*
GSM1184435	male	hip	Inf	1	03/12/2008
GSM1184436	female	knee	Inf	1	12/12/2008
GSM1184437	female	hip	Inf	1	12/12/2008
GSM1184438	female	finger	Inf	1	10/12/2008
GSM1184439	female	hip	Non-inf	1	03/12/2008
GSM1184440	female	knee	Non-inf	1	04/12/2008
GSM1184441	female	elbow	Non-inf	1	19/12/2008
GSM1184442	male	finger	Inf	1	04/12/2008
GSM1184443	female	Synovial Fluid Only	NA	1	12/12/2008
GSM1184444	female	hand	Inf	1	10/12/2008
GSM1184445	female	wrist	Inf	1	10/12/2008
GSM1184446	male	knee	Inf	1	12/12/2008
GSM1184447	female	NA	NA	1	12/12/2008
GSM1184448	female	NA	NA	1	03/12/2008
GSM1184449	female	hand	Inf	1	10/12/2008
GSM1184450	female	knee	Inf	1	10/12/2008
GSM1184451	female	hip	Non-inf	1	12/12/2008
GSM1184452	female	hip	Non-inf	1	10/12/2008

GSM1184453	male	knee	Inf	1	12/12/2008
GSM1184454	male	TKA	Non-inf	1	03/12/2008
GSM1184455	female	knee	NA	1	04/12/2008
GSM1184456	female	knee	Inf	1	03/12/2008
GSM1184457	female	NA	Inf	1	10/12/2008
GSM1184458	female	NA	Inf	1	04/12/2008
GSM1184459	female	knee	Inf	1	03/12/2008
GSM1184460	female	hand	Inf	1	12/12/2008
GSM1184461	female	foot	Inf	1	10/12/2008
GSM1184462	female	hand	Non-inf	1	12/12/2008
GSM1184463	female	hand	Non-inf	1	03/12/2008
GSM1184464	male	NA	NA	1	10/12/2008
GSM1184465	male	hand	Inf	1	12/12/2008
GSM1184466	male	hand	Non-inf	1	04/12/2008
GSM1184467	female	thumb	Inf	1	04/12/2008
GSM1184468	female	knee	Inf	1	04/12/2008
GSM1184469	female	hand	Inf	1	03/12/2008
GSM1184470	female	hand	Inf	1	12/12/2008
GSM1184471	female	NA	Non-inf	1	10/12/2008
GSM1184472	female	flexortenosynovium	NA	1	12/12/2008
GSM1184473	female	hand	Inf	1	03/12/2008
GSM1184474	female	hand	Inf	1	03/12/2008
GSM1184475	female	knee	Inf	1	04/12/2008
GSM1184476	female	wrist	Non-inf	1	04/12/2008
GSM1184477	female	hand	Non-inf	1	12/12/2008
GSM1184478	female	hand	Inf	1	12/12/2008
GSM1184479	female	wrist	Non-inf	1	12/12/2008
GSM1184480	female	knee	Non-inf	1	10/12/2008
GSM1184481	female	NA	Inf	1	12/12/2008
GSM1184482	female	NA	Non-inf	1	04/12/2008
GSM1184483	female	knee	Non-inf	1	03/12/2008
GSM1184484	unknown	unknown	unknown	2	05/11/2010
GSM1184485	unknown	unknown	unknown	2	09/11/2010
GSM1184486	unknown	unknown	unknown	2	03/11/2010
GSM1184487	unknown	unknown	unknown	2	05/11/2010
GSM1184488	unknown	unknown	unknown	2	05/11/2010
GSM1184489	unknown	unknown	unknown	2	04/11/2010

GSM1184490	unknown	unknown	unknown	2	03/11/2010
GSM1184491	unknown	unknown	unknown	2	04/11/2010
GSM1184492	unknown	unknown	unknown	2	05/11/2010
GSM1184493	unknown	unknown	unknown	2	05/11/2010
GSM1184494	unknown	unknown	unknown	2	04/11/2010
GSM1184495	unknown	unknown	unknown	2	04/11/2010
GSM1184496	unknown	unknown	unknown	2	09/11/2010
GSM1184497	unknown	unknown	unknown	2	03/11/2010
GSM1184498	unknown	unknown	unknown	2	05/11/2010
GSM1184499	unknown	unknown	unknown	2	09/11/2010
GSM1184500	unknown	unknown	unknown	2	05/11/2010
GSM1184501	unknown	unknown	unknown	2	09/11/2010
GSM1184502	unknown	unknown	unknown	2	03/11/2010
GSM1184503	unknown	unknown	unknown	2	09/11/2010
GSM1184504	unknown	unknown	unknown	2	05/11/2010
GSM1184505	unknown	unknown	unknown	2	04/11/2010
GSM1184506	unknown	unknown	unknown	2	04/11/2010
GSM1184507	unknown	unknown	unknown	2	05/11/2010
GSM1184508	unknown	unknown	unknown	2	09/11/2010
GSM1184509	unknown	unknown	unknown	2	09/11/2010
GSM1184510	unknown	unknown	unknown	2	09/11/2010
GSM1184511	unknown	unknown	unknown	2	03/11/2010
GSM1184512	unknown	unknown	unknown	2	05/11/2010
GSM1184513	unknown	unknown	unknown	2	17/11/2010
GSM1184514	unknown	unknown	unknown	2	04/11/2010
GSM1184515	unknown	unknown	unknown	2	04/11/2010
GSM1184516	unknown	unknown	unknown	2	05/11/2010
GSM1184517	unknown	unknown	unknown	2	05/11/2010

4.4.1.1. Attempting to incorporate healthy controls.

Healthy synovium samples were obtained from datasets GSE77298 and GSE82107, the healthy samples listed in both datasets are the same (as determined by comparing MD5sums of the .cel files), osteoarthritis is often used as a control; therefore this adds a total of 31 samples: 16 RA, 10 OA, and 7 no disease (ND). These samples have no additional metadata other than the disease state.

4.4.2. Hierarchical clustering of the samples.

After normalisation, samples were filtered to remove low-quality probes (displaying cross-hybridisation or some other deficiencies) as annotated by Affymetrix (HG-U133Plus_2.na36.annot.tsv). Following the removal of low-quality probes, the dataset was filtered further by requiring an Entrez ID – in the case of duplicates keeping the one with the highest variance – after filtering the dataset comprised of 17144 probes. For clustering, the top 40% more variable probes were selected (by standard deviation) resulting in 6858 probes being used.

Samples were clustered using 2 methods, initially using Ward's method on Euclidean distance of scaled and centred data, and then using Ward's Method on Euclidean distance on one minus Spearman correlation.

4.4.2.1. Testing for batch effects.

Quantifying the severity of the batch effects was done using the R package "BEClear"(146), genes were tested using the non-parametric Kolmogorov-Smirnov test (FDR adjusted) to identify those that are different as a result of batches (dataset and condition), after quantifying the genes that are different, they are binned relative to their difference from the median. To determine the severity of these effects, a weighted scoring based on these bins provides a Batch Effect Score (BEScore).

4.4.3. Investigating the immune component in the clusters.

As discussed in Section 3.1.5, 4263 probes representing markers of immune cells and synovial fibroblasts were utilised. After restricting the dataset to these probes, the data was transformed into z-scores before plotting.

4.4.4. Differential gene expression.

Differential gene expression for the clusters was determined by contrasting one cluster versus all clusters, using the limma package. Clusters were compared using a contrast matrix that follows the following expression ($X - \frac{x1+x2+x3}{3}$) where X represents the pathotype being compared and x represents the other clusters. The purpose of this contrast is that whilst it reduces power to detect effects by contrasting against itself, large group sizes would not outweigh small group sizes.

4.4.5. Pathology specific archetypal gene expression profile.

Pathology specific archetypical profiles were generated by taking the average for each cluster. Samples were then correlated against the archetype to show the similarity between a sample and representative pathology.

4.4.6. Identification of a predictive signature that differentiates the pathologies.

A predictive gene signature that discriminates the different pathologies was obtained by assessing several methods: Random Forest, Prediction Analysis for Microarrays (PAM), and Sparse Partial Least Squares Discriminatory Analysis (sPLS-DA).

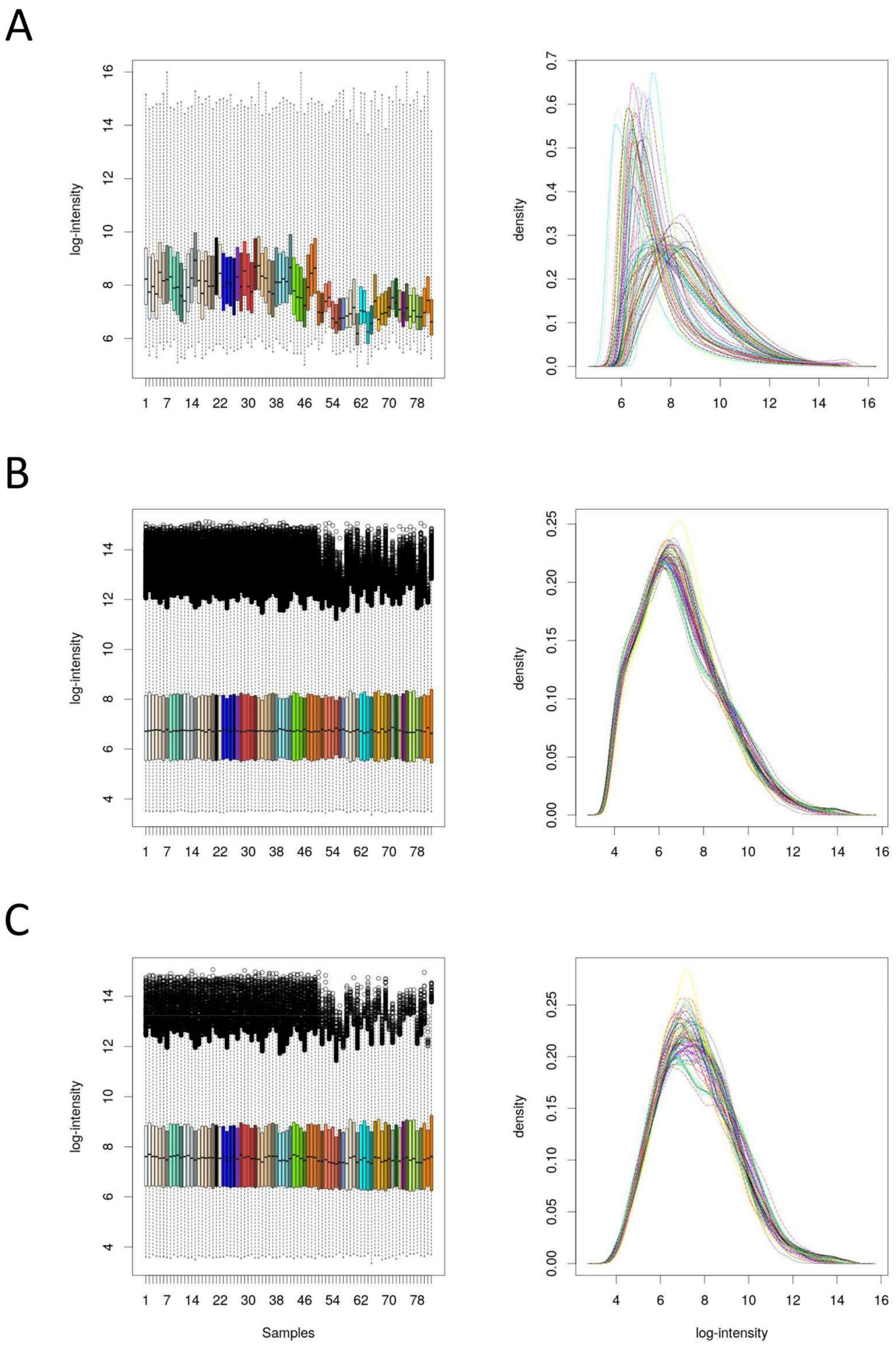


Figure 4.1: Initial assessment of sample distributions associated with the three datasets (GSE48780, GSE77298, and GSE82107).

These plots represent the distribution of synovial transcripts from patients with rheumatoid arthritis, osteoarthritis, and no disease. Transcript abundance is measured by probe fluorescence, intensity was log transformed to provide a useable scale. Boxes represent the 25th and 75th percentile, whilst centre band represents the median value, whiskers represent 1.5x the interquartile range, and outliers are indicated with circles. Density plots show the distribution of probe intensities across all the samples. ***A*** *shows the raw uncorrected data, as can be seen clearly from the boxplots samples from 50-82 have a different distribution of values, these samples are from datasets GSE77298 and GSE88107 which contain RA, OA, and no disease synovial samples.* ***B*** *shows the effects of RMA normalisation, bringing the samples into apparent normality across datasets.* ***C*** *represents probes from the normalised data that have been filtered in accordance to the methods utilised in the Dennis et al. paper (2014)*

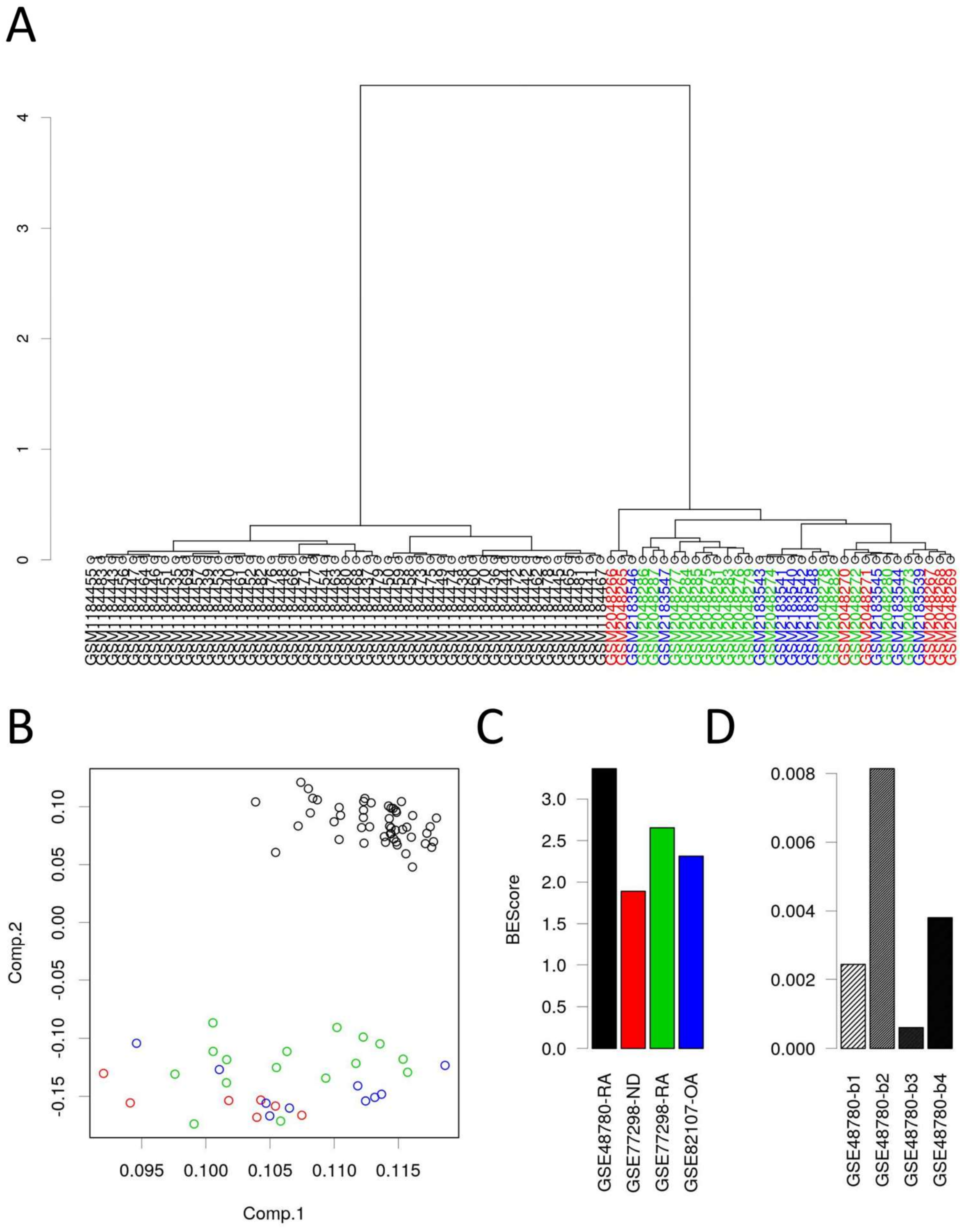


Figure 4.2: Detecting Batch effects in the data when attempting to include healthy controls.

Clear batch effects are visible in the clustering the samples from cohort 1 of GSE48780 with samples from GSE77298, and GSE82107. ***A*** *Samples were normalised with RMA, filtered to A class probes only and removing probes without an Entrez ID, with duplicate Entrez IDs being filtered to the probe with the highest variance. The top 40% most variable probes (n = 6858) were clustered (Wards method clustering, Euclidian distance on Spearman correlated data). GSE48780 (Black) contains only RA samples, GSE77298 contains RA samples (Green) and No Disease (Red), whilst GSE88107 contain OA samples (Blue). As can be seen clearly from the split in the dendrogram the difference between the datasets (GSE48780 vs GSE77298 & GSE88107) is far bigger than the difference between diseases.* ***B*** *PCA on all probes shows the same pattern of discrimination between the datasets.*

In order to quantify the severity of batch effects, the R package "BEClear" was utilised. ***C*** *After identifying all of the genes that are different between batches (dataset and condition) using Kolmogorov-Smirnov test (p-values adjusted using FDR), Batch Effect Scores (BEScore) that determine the severity of the batch effect was determined with a weighed score based on bins assigned by differences to median.* ***D*** *To provide context for the score, random batches were assigned to samples in cohort 1 of GSE 48780, clearly visible is how small the BEScore is relative to* ***C****.*

4.5. Results.

4.5.1. Batch effects.

As GSE48780 lacked any control group, comprising entirely of RA patients, healthy controls were incorporated into the analysis using samples from GSE77298 and GSE82107. As these transcriptomic datasets have been obtained from different sources, it was important to investigate, and (if present) control for batch effects.

Looking at the raw uncorrected data Figure 4.1A there is a fundamental difference in the signal from the microarray. In the boxplots, there is a distinct change in the interquartile ranges (inside the boxes) from sample 50 onwards that reflect the samples from GSE77298 and GSE82107, and this is further typified by the density plots, where there are two different distributions of probe intensities.

After RMA normalisation (Figure 4.1B), it does appear that the samples are normalised correctly, with the interquartile ranges being consistent across the samples, though outliers still show different behaviour.

After filtering probes as detailed in Section 4.4.2, this difference in the outliers for the other samples becomes more pronounced and causes changes to the distribution of probes.

Figure 4.2A shows the results of unsupervised clustering of the samples according to the adapted methodology discussed in Section 4.4.2 identifies two primary clusters. Identifying these samples in these clusters segregates the GSE48780 samples from the samples of GSE77298 and GSE82107. Given that the latter two datasets contain RA, OA, and ND, therefore, it is clear that a batch effect that separates these datasets rather than differences between diseases. This is further exemplified in Principle Component Analysis (PCA) as illustrated in Figure 4.2B, which utilises all of the probes, and shows the same discrimination between the datasets rather than by condition.

Figure 4.2C shows a high BEScore illustrating that there is a fundamental difference between these datasets, as this score is essentially unitless, a comparison can be made with Figure 4.2D which is the result of a random assignment of batches within cohort 1 samples clearly showing how strong this effect is.

As discussed in Chapter 3, methods that have been developed to address batch effects also have the habit of destroying biological variance. Given this limitation, it was not possible to integrate these healthy samples as controls for further analysis. As these datasets (GSE77298 & GSE82107) only have a small number of RA samples, and no information with regards pathology, it is of limited value to explore the biology of the disease. Therefore, going forward the data analysed is from GSE48780 exclusively.

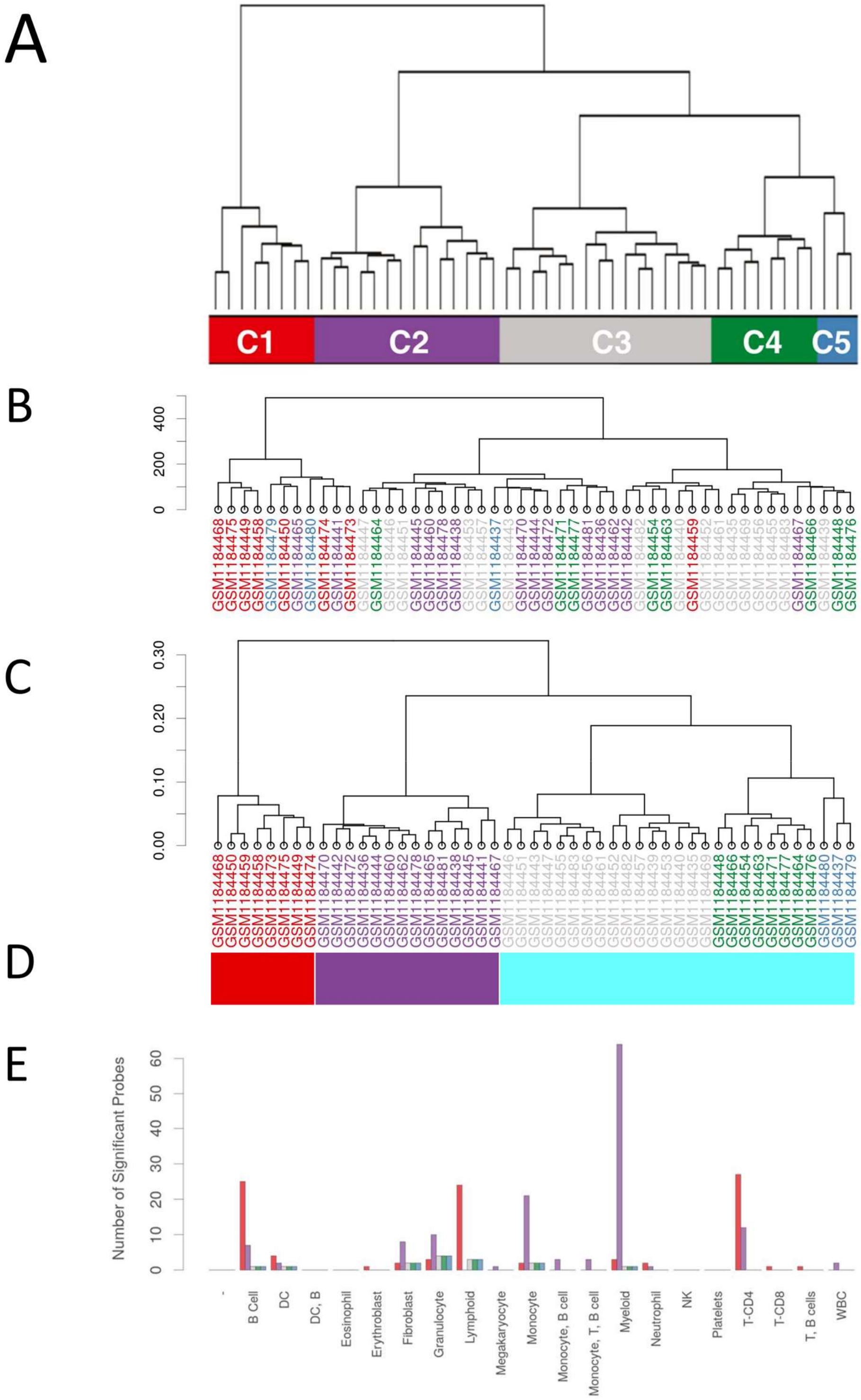


Figure 4.3: Unsupervised clustering of the training dataset reveals the three pathologies of rheumatoid arthritis.

Replicating the hierarchical clustering of the samples as performed in the Dennis et al. ***A*** *represents the original clustering from the paper (adapted from figure 1A). For both figures* ***B*** *and* ***C*** *samples were filtered on the following* *parameters: exclusion of non A-class probes (as determined by Affymetrix quality assessment); exclusion of any probes missing Entrez ID; and restrict any probes with multiple Entrez ID's were restricted to the one with the highest variance; probes were then ranked by variance, and the top 40% most variable probes (n = 6858).* ***B*** *replicates the methodology in the Dennis et al paper, with the samples scaled and centred, before clustering (Ward's method on Euclidian distance). However, the number of samples belonging to each cluster does not match the clustering seen in the original paper. Colours reflect clusters as identified within* ***C***. ***C*** *is a reproduction of the dendrogram of the original paper, with a slight modification of the methodology. Samples are clustered in a similar manner but utilise a 1 – correlation between samples to calculate the Euclidean distance before using Ward's method. Due to small differences possible with revisions to Affymetrix's quality assessment, there are minor differences in the tree structure. However,* ***C*** *reproduces the grouping of 8 samples (red), 14 samples (purple), 16 samples (grey), 8 samples (green), and 3 samples (blue) that represent C1 through 5 respectively in the original figure. Clinically only 3 pathologies are described, and these are mapped in* ***D****, showing Follicular (Red) and Diffuse (Purple) remains the same, however the three clusters (Low Inflammatory, Fibroid, and C5) represent a similar immunological profile (****E****) and are therefore considered to be Pauci Immune (Cyan).* ***E*** *Looking at the significantly upregulated (adjusted p-value ≤ 0.05, ≥1.5 fold change) immune markers (see chapter 4.5.3), we see that there is no difference in the upregulated marker probes between Low Inflammatory, Fibroid, and C5.*

4.5.2. Reproduction of the original clustering.

One important factor missing from the metadata is the synovial pathology that these patients exhibit, therefore it is needed to reproduce the clustering to identify the samples.

Figure 4.3A which is derived from the original paper shows 5 clusters, labelled C1-C5, with 8, 14, 16, 8, and 3 samples to the clusters, respectively. Following the methodology exactly as laid out in the original paper results in the dendrogram shown in Figure 4.3B, immediately evident is that the dendrogram has no relation to that shown in Figure 4.3A. The colouring shown on this tree indicates the classification of the samples as determined with the reproduction seen in Figure 4.3C.

Reproducing the clustering required a modification of the methodology, instead of using scaled and centred data it utilises Spearman correlation of the samples. This clustering reproduces the same overall shape of the dendrogram, with very small differences that may be the result of different quality assessment annotation – as the version used is not stated. Cutting the tree to provide five clusters as previously identified results in the same number of samples seen in the original, and these samples have been coloured to match.

However, as previously discussed there is only histological evidence for 3 pathologies, and when investigating the immune profile of the samples the clusters previously identified as C3—C5 (see Figure 4.3A) have a similar immune profile. Looking at the number of significantly upregulated immune markers (Figure 4.3E) demonstrates that these clusters are behaving the same, whilst C1 and C2 demonstrate very different profiles. Therefore, taking this lack of difference between these three clusters and the histological evidence for 3 pathologies, they were collapsed into a single category as indicated in Figure 4.3D.

4.5.3. Immune component expression.

To validate that the clusters identified reflect real pathologies, the data was restricted to probes that can be utilised as markers that discriminate the different immune cells(122–

125,147) or synovial fibroblasts from RA(126) (see Section 3.1.5 for an explanation of how these were derived).

Figure 4.4A shows the scaled expression data for the dataset ordered by the cell type markers. In general, there is a general enrichment of expression for the immune cells in the follicular and diffuse clusters, with pauci-immune samples exhibiting a generally downregulated expression across the marker list. Unexpectedly, the fibroblast markers do not show a consistent upregulation across the samples, however, more of the fibroblast markers are significantly upregulated when looking at differentially expressed genes (Figure 4.4B). Figure 4.4B demonstrates the expected immune profile associated with the three pathologies when we look at significantly upregulated genes within this marker list: Follicular (red) being enriched for B cells and general lymphoid markers, as well as highest levels of T cells; Diffuse (purple) being massively enriched for monocyte and myeloid markers; and Pauci-immune (cyan) showing little enrichment for the immune cells.

Clustering the rows in Figure 4.5 illustrates 5 clear blocks of gene expression, however, identifying the individual cell types here would be almost impossible – therefore the annotation has been collapsed down to lineages. This annotation demonstrates clear enrichment of specific lineages within these clusters, notably lymphoid lineage markers in cluster 5 and myeloid lineage markers in clusters 1 and 2

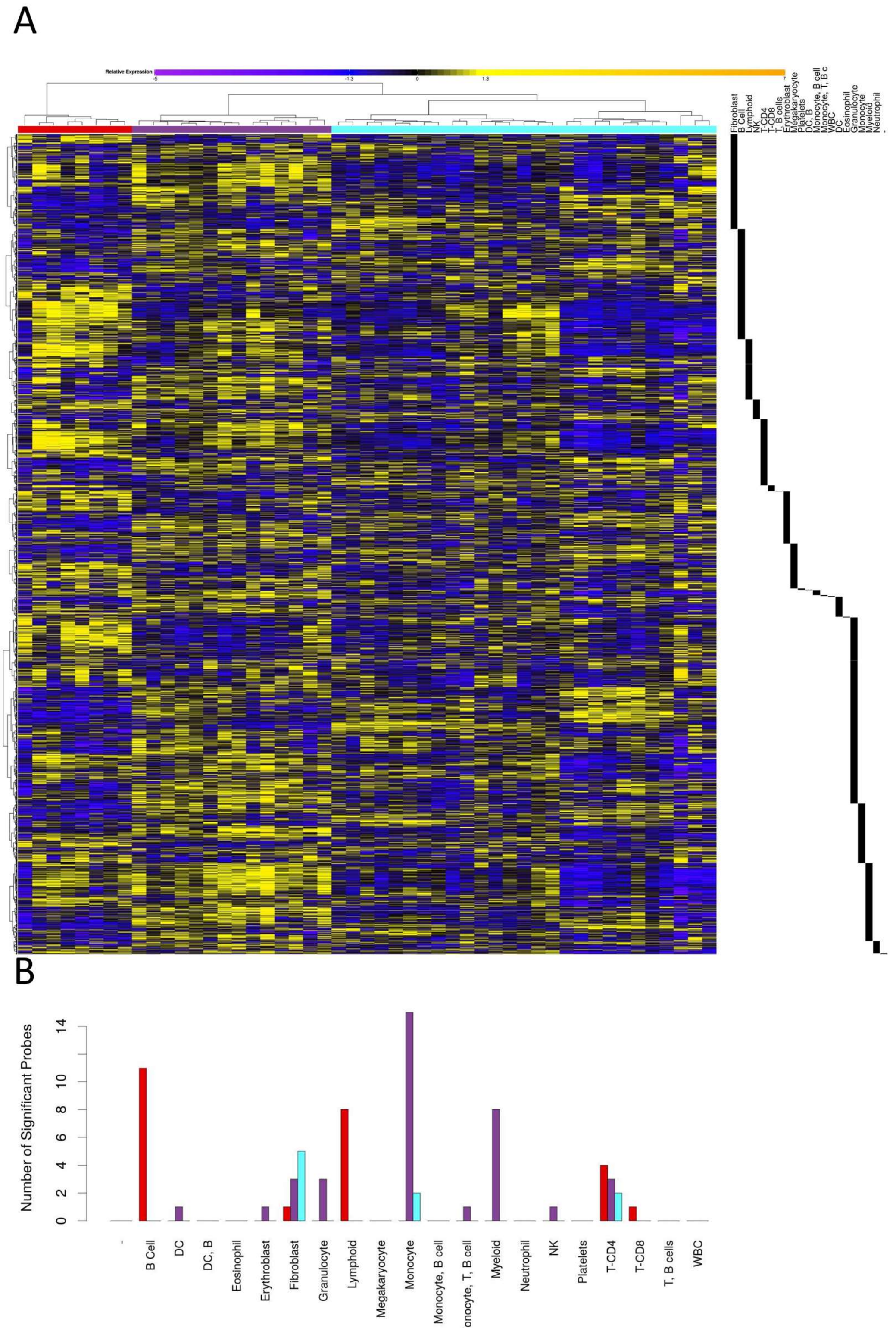


Figure 4.4: Immune profile of clustered synovitis pathologies.

Looking at the scaled expression of 4265 marker genes across the cohort 1 of GSE48780. Immune markers were derived from five public databases used in the R package "CellMix", using the Abbas et al 2005 (IRIS), Palmer et al 2006, Abbas et al 2009, Watkins et al 2009 (HaemAtlas), and Grigoryev et al 2010 marker lists. These were merged, with unique or matching annotations used, and where markers identified differing cell types the closes lineage precursor was used. Fibroblast markers were derived from Filer et al 2015 which identified markers that were uniquely upregulated in synovial fibroblasts in RA patients compared to skin, bone, or synovial derived cells from RA and OA patients. ***A*** *shows the increased expression (yellow-orange) of immune genes reflective of the immune infiltrates observed in histological characterisation of the pathologies. Marker genes are indicated by the* *annotation bars on the right of the figure. For example an increased expression of B-cell markers in the Lymphoid, Granulocyte markers in the myeloid, and a general absence in the Pauci immune. This enrichment is illustrated in* ***B****, with significantly upregulated markers (≥1.5 fold and p ≤ 0.05 (Bonferroni corrected)). Follicular samples exhibit an enrichment for B, T and general lymphoid cells, whilst Diffuse samples are enriched for monocytes, granulocytes and myeloid cells, and Pauci Immune is slightly enriched for fibroblast markers.*

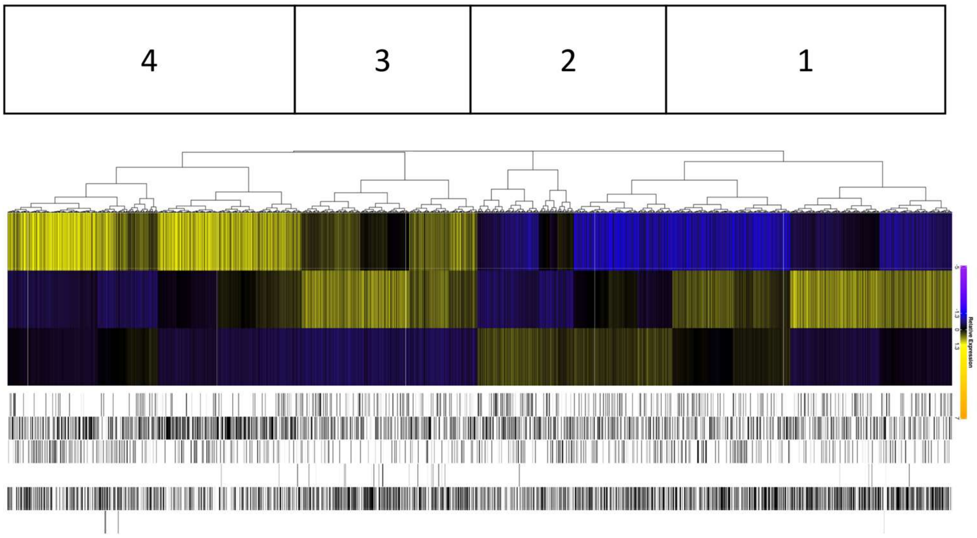


Figure 4.5: Clustering the rows identifies blocks of genes with similar behaviours.

Four patterns of expressions are seen, for example up in Diffuse relative to Follicular and Pauci (pattern 1). As this clustering reorders the rows, this makes it impossible to visualise relative to cell type, and therefore annotation has been restricted to cell lineage. Then when looking at the patterns we can see that pattern one has considerably more myeloid lineage markers than pattern Four, which itself is enriched for lymphoid markers.

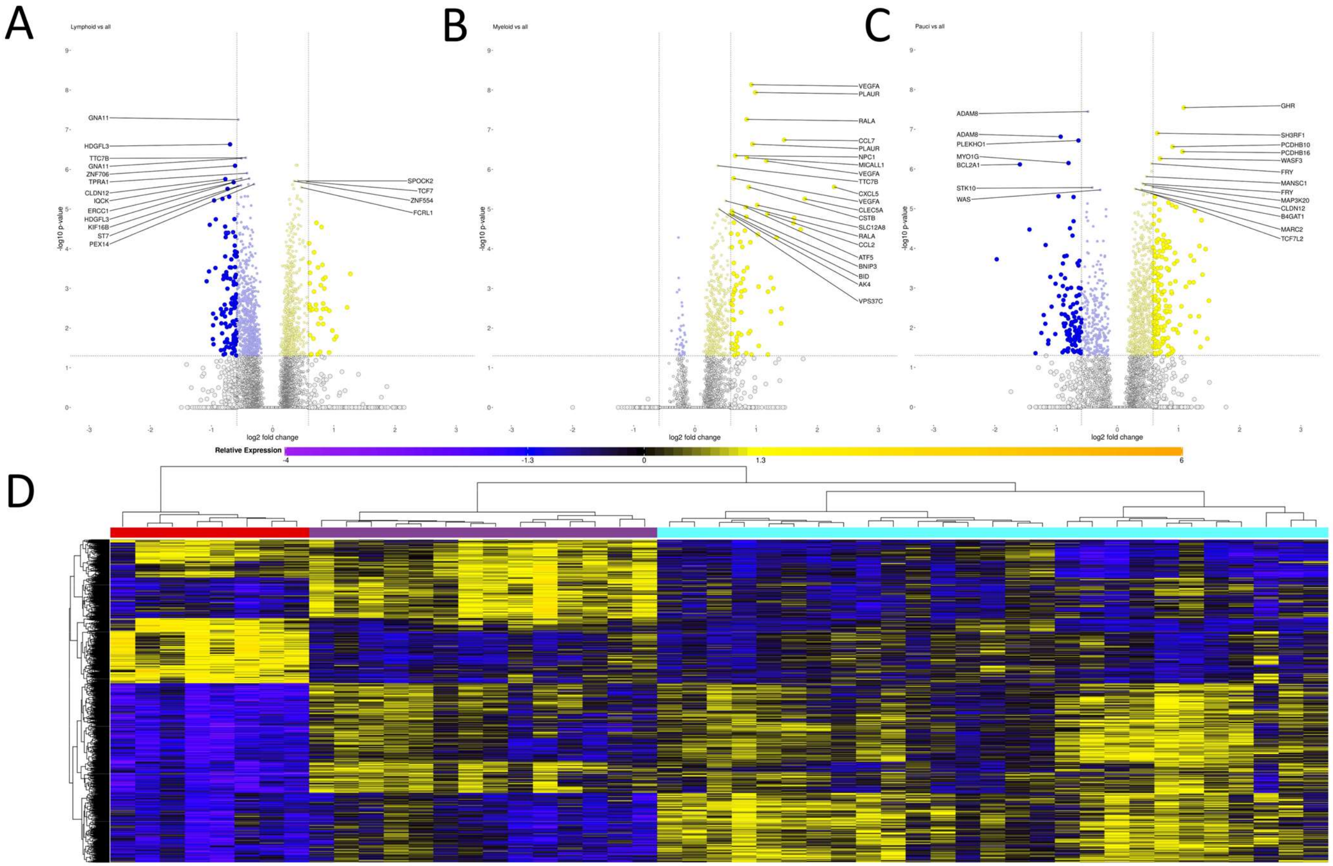


Figure 4.6: Differentially expressed genes between the pathologies.

***A**, **B**, & **C** show volcano plots contrasting the individual pathotypes (Follicular, Diffuse, Pauci immune respectively) against all pathologies combined, x-axis represent fold change, whilst y-axis shows the –log10(p-adjusted) (Bonferroni corrected). This contrast was necessitated by the lack of a control group but allows all results to be directly compared against each other. Given the reciprocal nature of this contrast, further exploration into what pathways focusses on the upregulated genes that define what is increased in that pathology, rather than downregulated genes that are more of a representation of upregulation in another pathology. **D** Illustrating all the significantly differentially expressed genes across the pathologies (p.adjusted ≤ 0.05).*

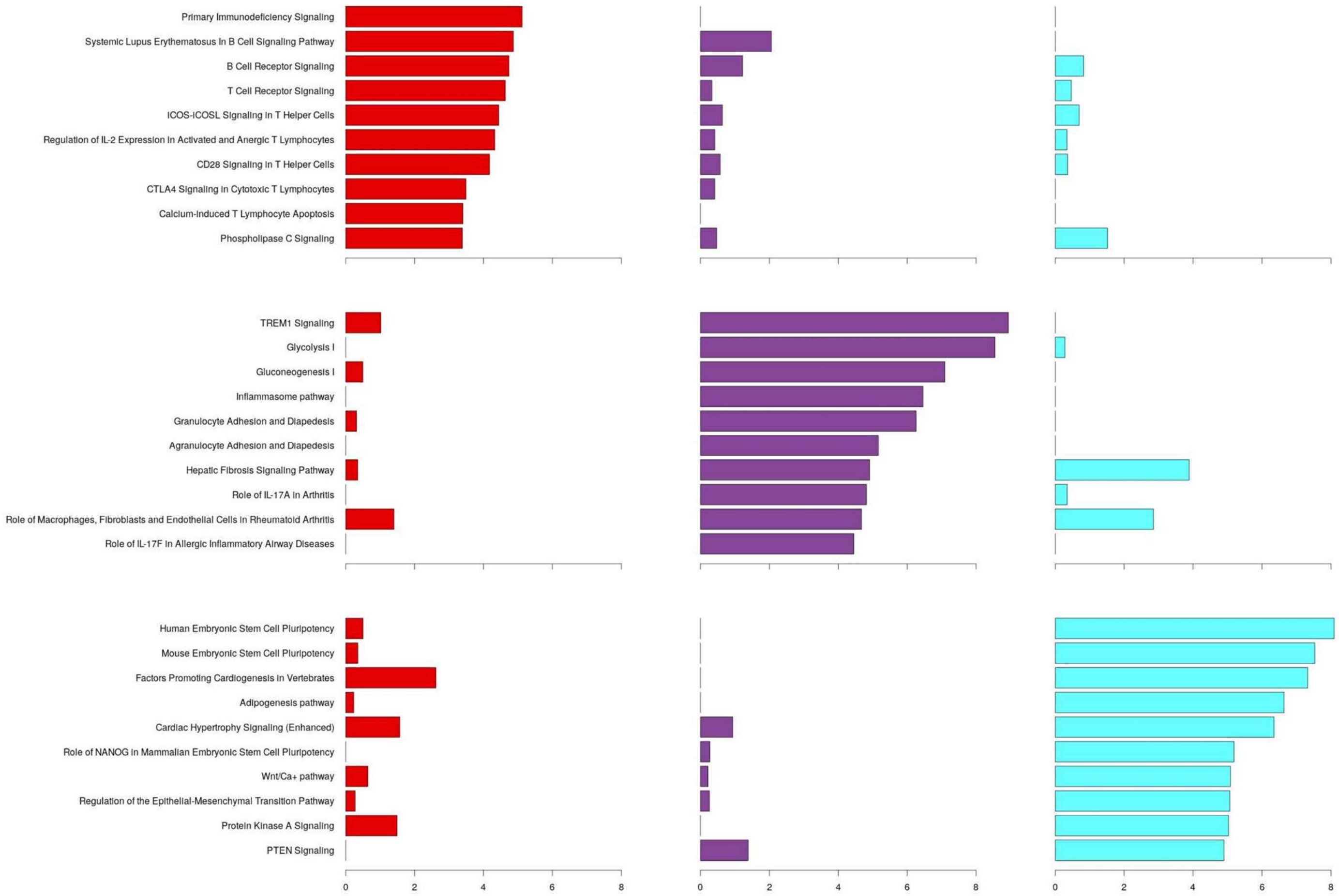


Figure 4.7: Investigating the upregulated pathways in the three pathologies using enrichment analysis in Ingenuity Pathway Analysis (IPA).

Probes were selected after differential expression, retaining only those that were significant (≤0.05) after adjusting using Bonferroni correction, and had a fold change greater than 0 (therefore only upregulated). These were then subjected to the core analysis for each pathology in IPA, before making a comparative analysis, exporting, and plotting the top 10 pathways in each pathology.

4.5.4. Differential expression.

To investigate the pathways associated with the pathologies requires identification of differentially expressed genes, however, as GSE48780 lacks a control group, this necessitated comparing one group vs all groups. Unfortunately, this does result in the fact that downregulated genes in one group are also upregulated genes in another, therefore downstream analysis focusses primarily on upregulated genes in each group. The positive of this design, however, is that consequently, all comparisons are directly comparable between all three clusters, rather than being done in a pairwise manner and accordingly more tests. As these have been characterised by the immunophenotypic expression discussed in section 4.5.3, clusters will henceforth be referred to by the name of the representative synovial pathology.

Figure 4.6 shows the differentially expressed genes across the 3 pathologies, notably, as previously discussed, the reciprocal nature of differentially expressed genes given this method of comparison. 2428 probes are differentially expressed across the three pathologies at a significance level of less than or equal to 0.05 after Bonferroni correction, reflecting 2016 genes that are significantly different. This breaks down to 1188 probes or 1068 genes in Follicular, 453 probes and 373 genes in Diffuse, and 1159 probes and 930 genes in Pauci-immune. However, given this reciprocal nature, further analysis of differential expression was restricted to only those genes that were upregulated.

Figure 4.7 illustrates the significantly upregulated pathways in the three pathologies, the follicular pathology characterised by lymphocyte-associated pathways, diffuse with myeloid associated pathways and metabolism, and pauci-immune with tissue remodelling.

Significant genes upregulated in the follicular pathology highlight the role of the adaptive immune system, with numerous genes associated with T-cell receptor signalling (*ZAP70, CTLA4, ICOS, LAT, CD247*) and B-cell signalling (*CD7B, PRKCB, CARD11, CD22*), as well as the differentiation and proliferation of lymphocytes. Multiple cytokine signalling pathways are also identified (*IL-2, -3, -4, -5, -7, -9, TNFα*) and therefore downstream activation of the PI3K-Akt (*TCL1A, CCND3, CCNE1-3, COL4A3, PKN2*), NF-κB (*TNFSF14, TNFRSF13C, LTB*) and Jak-STAT signalling (*STAT5B, IL2RB, IL21R, IL12RB1*).

Significant genes upregulated in diffuse shows alterations to metabolism in particular gluconeogenesis and glycolysis (*GPI, TPI1, PGAM1, ENO1-2, GAPDH*), pentose phosphate (*TANDO1, PGLS, ALDOA*) and Fructose Mannose metabolism (*SORD, HK2*). Induction of innate immune responses can be seen in the upregulation of genes associated with TLR (*CXCL8,* *MAP2k3, TICAM1, MYD88, IFNAR1*), TREM1 (), and Fibrin Complement Receptor 3 Signalling Pathways (*SRC, CCL2, CXCL3*).

Significant genes upregulated in pauci-immune demonstrate a role in controlling tissue remodelling through TGF-β regulation of the extracellular matrix (*SYNM, FBLN5, TGFB2, BMP4*), VEGF signalling (*PPP3CA, PIK3CA, PIK3R1, PTK2*), and wnt signalling (*TCF7L1-2, PLCB4, RYK, FZD7-8:10*).

4.5.5. An archetype for each pathology.

To provide a reference point for what each pathology "looks like", an archetypical expression profile was constructed by taking the average expression for all probes for each of the representative clusters. Comparison of individual samples was done by correlating the entire transcriptome (all 54675 probes) against the three pathologies. Whilst focused on the immune markers the pathologies Illustrate a clear difference in expression, across the entire transcriptome the samples are remarkably similar, necessitating a correlation scale of 0.97-1 to visualise the differences between the samples and archetype (Figure 4.8).

Obviously, the archetypes being derived from these samples, they exhibit a strong correlation for each of the representative samples; however, it is notable that there is considerable crossover between some of the pathotypes. Examining the follicular pathology, it has a relatively weak correlation with samples from the other pathologies, whilst diffuse and Pauci show considerable overlap.

This archetypical signature can be compared against other datasets (see Chapter 5) on the same (or similar) microarray platform but presents challenges when looking at other transcriptomic platforms and organisms due to mapping multiple probes to a single gene or homologous genes in different organisms.

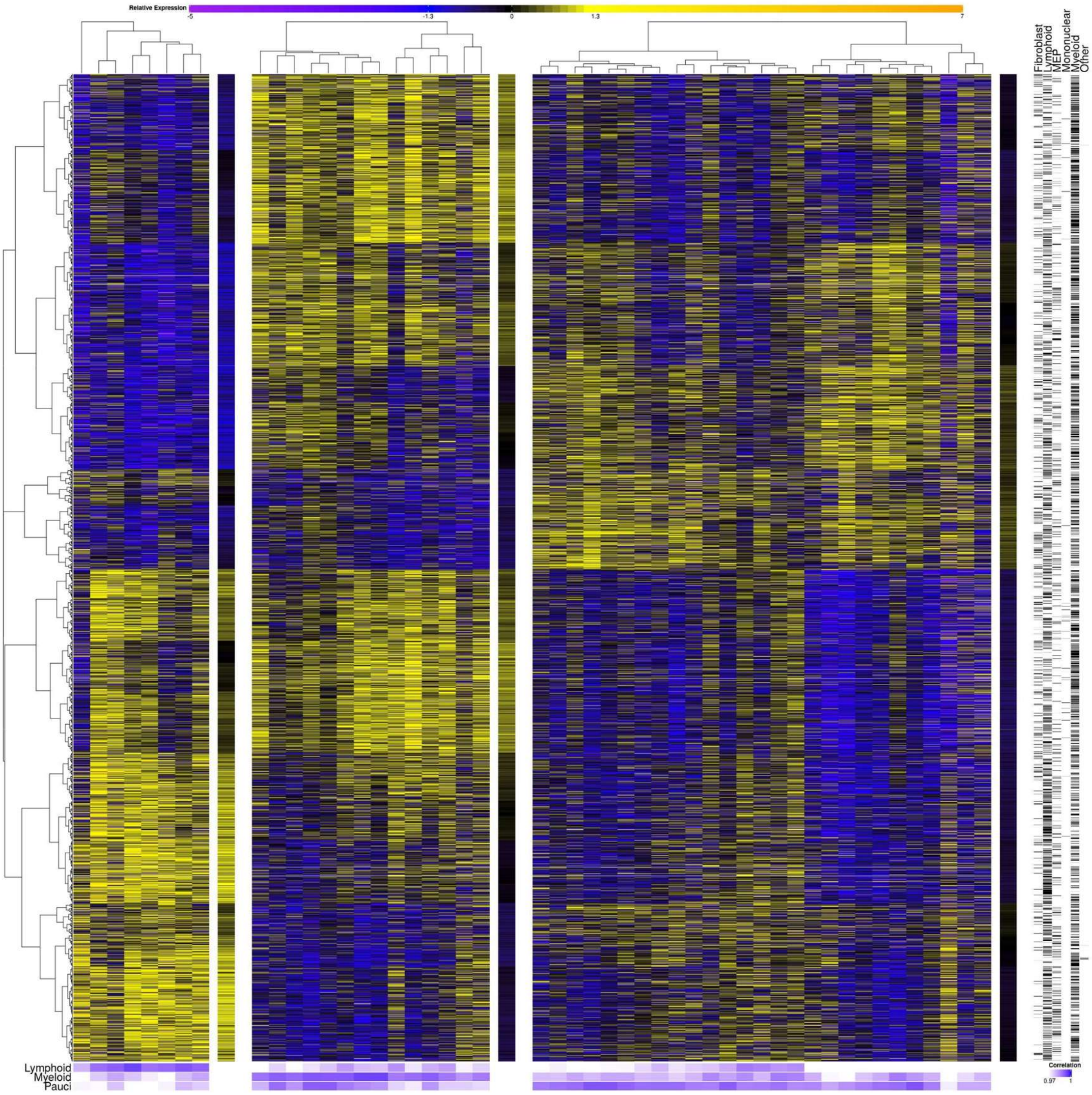


Figure 4.8: Identifying the archetypical expression for the different pathologies.

Archetypes were constructed by taking the average expression for every probe for the different clusters. By correlating the entire transcriptome of the individual samples against the architypes provides a Pearson's correlation coefficient for every sample, the stronger the correlation the darker the shade of blue.

4.5.6. Gene signature.

Whilst interrogation of the entire transcriptome allows stratification of the patients, this isn't the most efficient use of resources to determine the pathology. Therefore, a small gene signature that allows the stratification of patients would provide a valuable resource for determining the optimal therapeutic choices.

To approach this the first method utilised random forests to select variables that segregate the pathologies. After normalisation, the entire dataset was filtered to require that probes exhibit a minimum expression in a proportion of the samples, in this case, a minimum expression of 100 in 20% of samples, this reduced the number of probes in the dataset from 54'675 down to 37'188. Using the pathotypes as defined in Section 4.5.3, an initial random forest of 100'000 iterations was performed, using all of these probes demonstrates very poor performance in classifying the follicular pathology (AUC: 0.558) (Figure 4.9A. Figure 4.9B illustrates the importance of the variables in the trees a cut-off of 25 genes was used which is illustrated in Figure 4.9C. Reducing the number of variables improves the AUC for diffuse and pauci, but follicular still exhibits poor classification rates (Figure 4.9D). This is reflected in the clustering, which fails to partition the samples according to their pathologies (Figure 4.9E)

The second method utilised was using the package Prediction Analysis for Microarrays (PAMr) which utilises the nearest shrunken centroid method to identify candidate genes for classification. Using the default 10-fold cross-validation it is possible to determine the threshold with minimal misclassification errors and the least number of genes. Figure 4.10A shows the number of genes at a given threshold, whilst the misclassification rate is shown in Figure 4.10B, as the misclassification rate started increasing for follicular samples at a value of 4.42, this threshold was utilised going forward. Given this threshold, 158 probes are employed in the classifier (Figure 4.11), the cross-validated probabilities for these demonstrates the performance of the classifier (Figure 4.12A), as well as the unsupervised hierarchical clustering (Figure 4.12B), and the Receiver Operator Characteristics (Figure 4.12C).

The third method utilised was sparse Partial Least Squares Discriminatory Analysis (sPLS-DA)(148). To determine the appropriate number of elements for each component, bootstrapping using the related bootsPLS(14) package was employed to calculate the optimal numbers. After 997 iterations, fit models were extracted from the data. This is a computationally expensive analysis, with the 997 iterations taking over 24 hours to run across 2 nodes of the Hawk supercomputer (2x 20 cores per node). Prior to this undertaking, a pilot study was performed with a lower level of cross-validation (5x) and fewer iterations prior to implementation of multithreading necessitating around 18 hours of computation to generate 15 iterations. As there are a large number of observations going in, the number of iterations is potentially not high enough to stabilise the selection of predictors, therefore there is some variability to probes utilised in the model. To account for this variability, one hundred random seeds were created, and a fit model performed using these seeds. Subsequently, the performance of the model evaluated using a comparison of hierarchical clustering. To account for this multiple sampling, identification of any common traits between the models was also examined. Additionally, a signature derived from a prior pilot study investigating the technique was utilised.

- Performance of the models was evaluated with two methods, an initial screening based on its ability to cluster the groups as assessed by a tanglegram, looking for complete clustering of pathologies – belonging to distinct branches with minimal misclassification.

- And those models with the best performance then explored using the Receiver Operator Characteristics (ROC).

Figure 4.13A illustrates 2 good fits (fit1556 and fit8945) which whilst not perfectly stratifying the samples retain a hierarchical separation of the three pathologies. Whereas Figure 4.13B has multiple samples miss-clustered, and the hierarchical organisation of the pathologies is destroyed.

Within the seeds, a number of models exhibited good performance in clustering and area under the curve, as illustrated in Appendix Figure 10.1. Table 4.1 shows the overlap between the probes in these good models, as well as a "bad fit", looking across these multiple models identify a core set of genes that are common to all. The similarity of multiple models at the 28/29 probes across multiple seeds, illustrated in Table 4.1, demonstrates how close the models are to being stable, but due to the computational cost, no further iterations were performed.

A core signature was derived from all of the fit models, resulting in 17 probes that are consistently present. Moreover, 14 probes were consistently present in the first component, and 3 more in the second (Figure 4.14C & 4.14 D). This signature performs well when looking at the area under the receiver operator curve (AUROC) (Figure 4.14B), and separates well in PCA space (Figure 4.14xE). However, this is not reflected in the unsupervised clustering, with some misclassification and non-distinct clusters, as shown in Figure 4.14F but ultimately retaining hierarchical order associated with pathologies.

A pilot study into the technique, using a low iteration implementation (15 iterations), identified 14 probes that stratified the pathologies. As two of the probes (214435_x_at & 224880_at) are for the same gene, RALA, the latter probe was removed with no major effects on the clustering or performance, resulting in a 13 probe/gene signature going forward. Investigating this signature in the larger iteration dataset demonstrates a good classification once the second component is included, as assessed by the ROC (Figure 4.15A & 4.15B). As with the other sPLS-DA results, this signature shows good separation in the PCA space (Figure 4.15E) and shows good performance in the unsupervised clustering (Figure 4.15F).

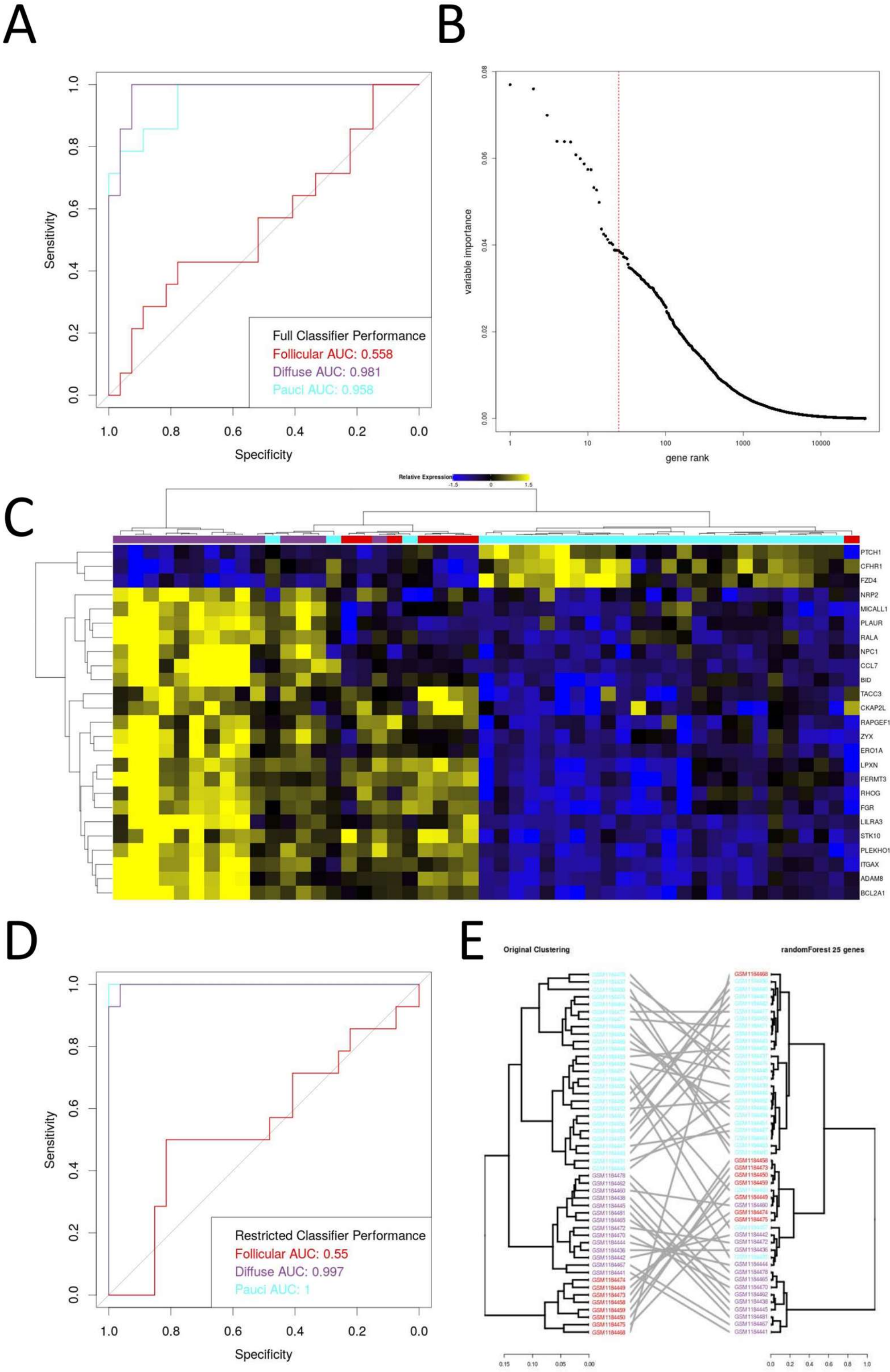


Figure 4.9: Predicting pathologies using random forest:

Probes were filtered to require a minimum expression of 100 across 20% of probes resulting in 37 thousand probes going into the classifier. ***A*** *Classifier performance using the full dataset, demonstrating relatively good performance for Pauci and Diffuse, but barely predicting Follicular better than Random.* ***B*** *Plotting the importance of genes in describing variance, red line shows the intersect for 25 probes that were utilised in the restricted classifier.* ***C*** *Heatmap of the scaled expression of the 25 probes, (**D**) restricting the dataset going in improves classifier performance for Pauci and Diffuse, but Follicular still exhibits poor classification.* ***E*** *This is reflected in the unsupervised clustering which shows mixed clustering.*

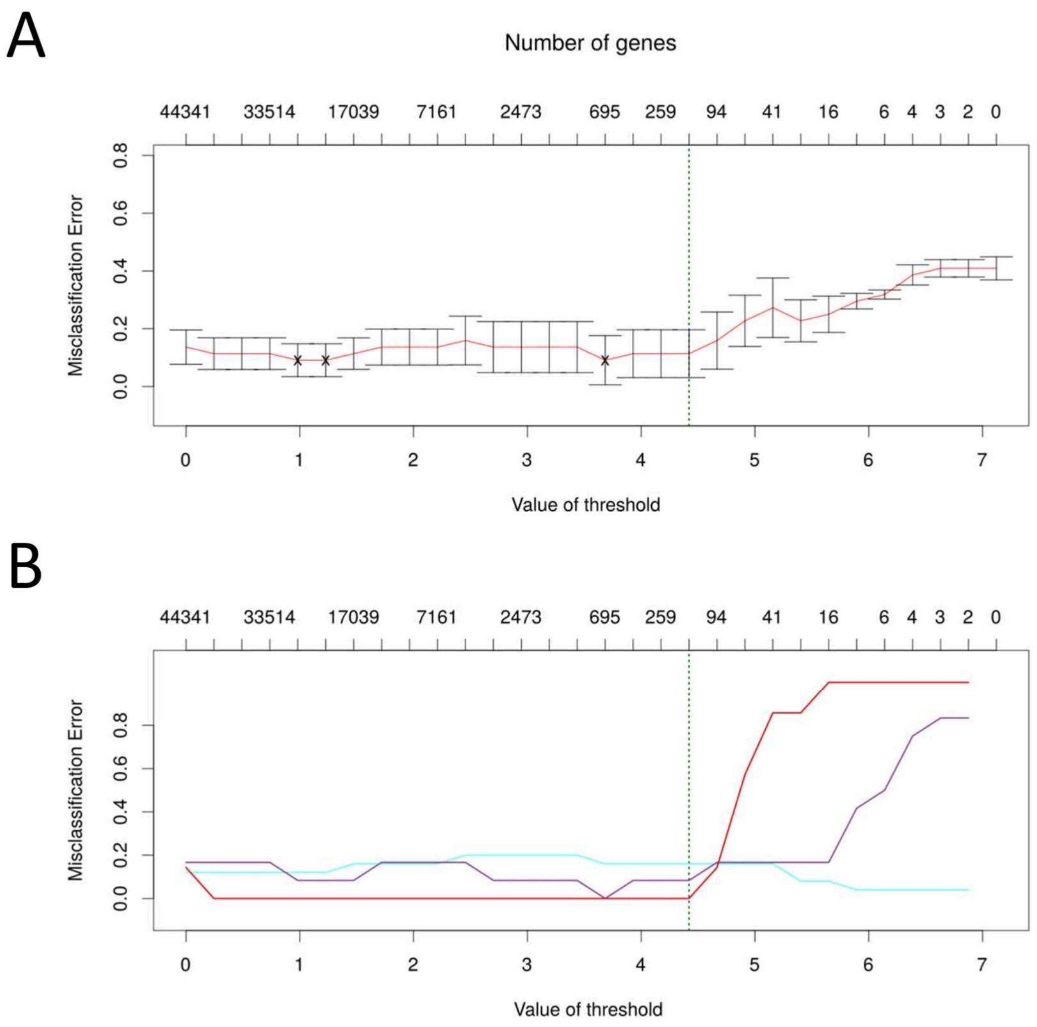


Figure 4.10: Identification of gene signature that discriminates pathologies using the Prediction Analysis for Microarrays (pamr) package .

***A** Plotting the error rate to determine the cut-off threshold that balances minimisation of the misclassification error and the number of genes needed in predicting pathology, **B** shows the pathology specific error rates, as well as the number of probes included at a given threshold level. The threshold value (green dotted line) of 4.42 was chosen as this is the point where misclassification rates for the follicular pathology starts to increase, and this minimises the number of genes whilst keeping the classification accuracy.*

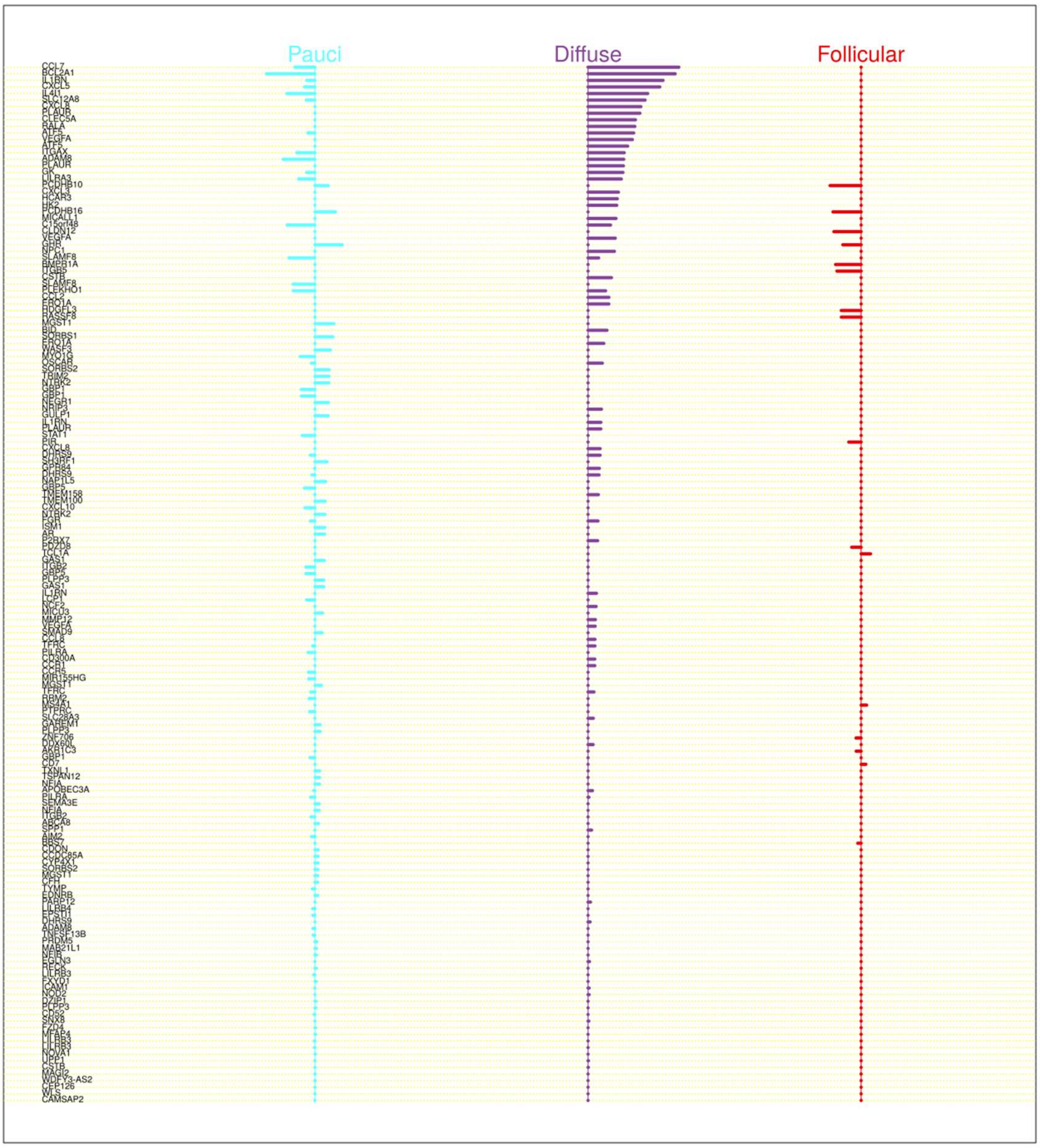


Figure 4.11: The contribution of the 158 probes in discriminating the different pathologies at a threshold value of 4.42.

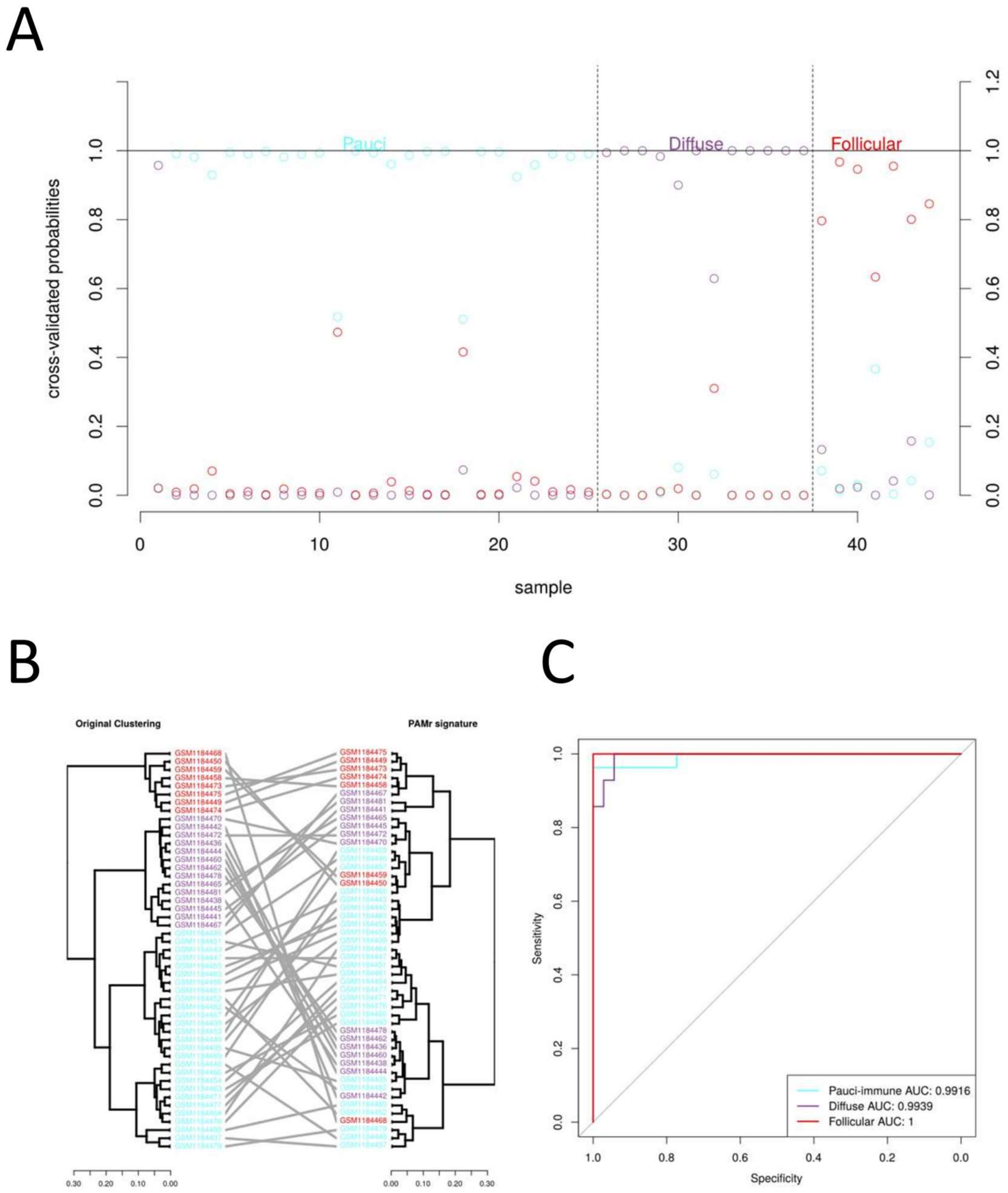


Figure 4.12: The performance of the 158 probe signature derived through the Prediction Analysis for Microarrays (PAMr) package.

***A** The cross-validated (10 fold) probabilities for pathology prediction. **B** Illustrating the clustering obtained using these 158 probes compared to the clustering obtained in Figure 1.3. **C** The Receiver Operator Characteristics for classification performances at the 4.42 threshold.*

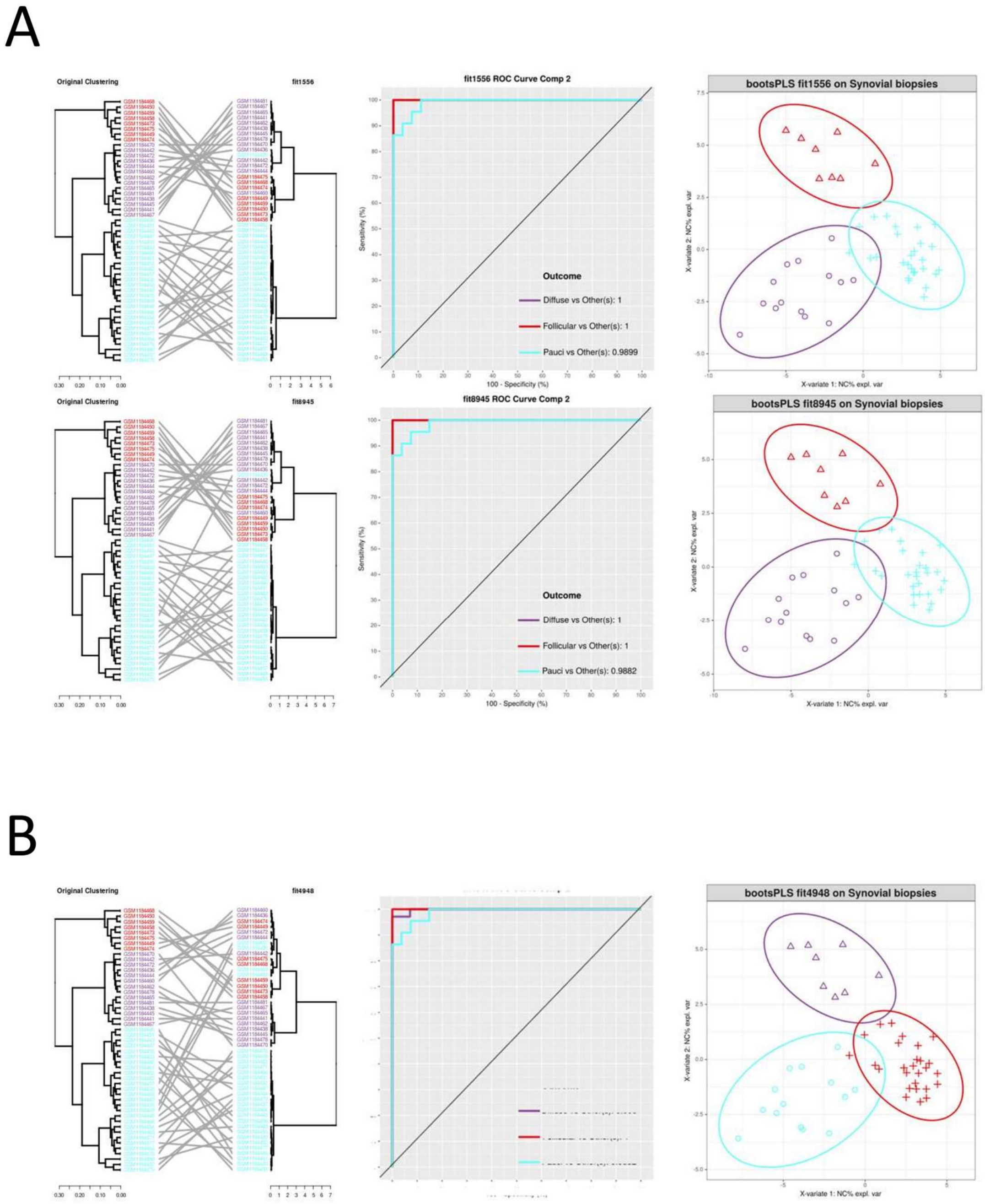


Figure 4.13: Assessing fit performance using unsupervised clustering based on the gene signatures identified by bootsPLS

*Whilst all fits appear to perform well when looking at the ROC performance and the principle component separation. Unsupervised clustering demonstrates that some fit models retain a hierarchical order associated with pathology (**A**), some models lose this order, and are therefore considered bad fits (**B**)*

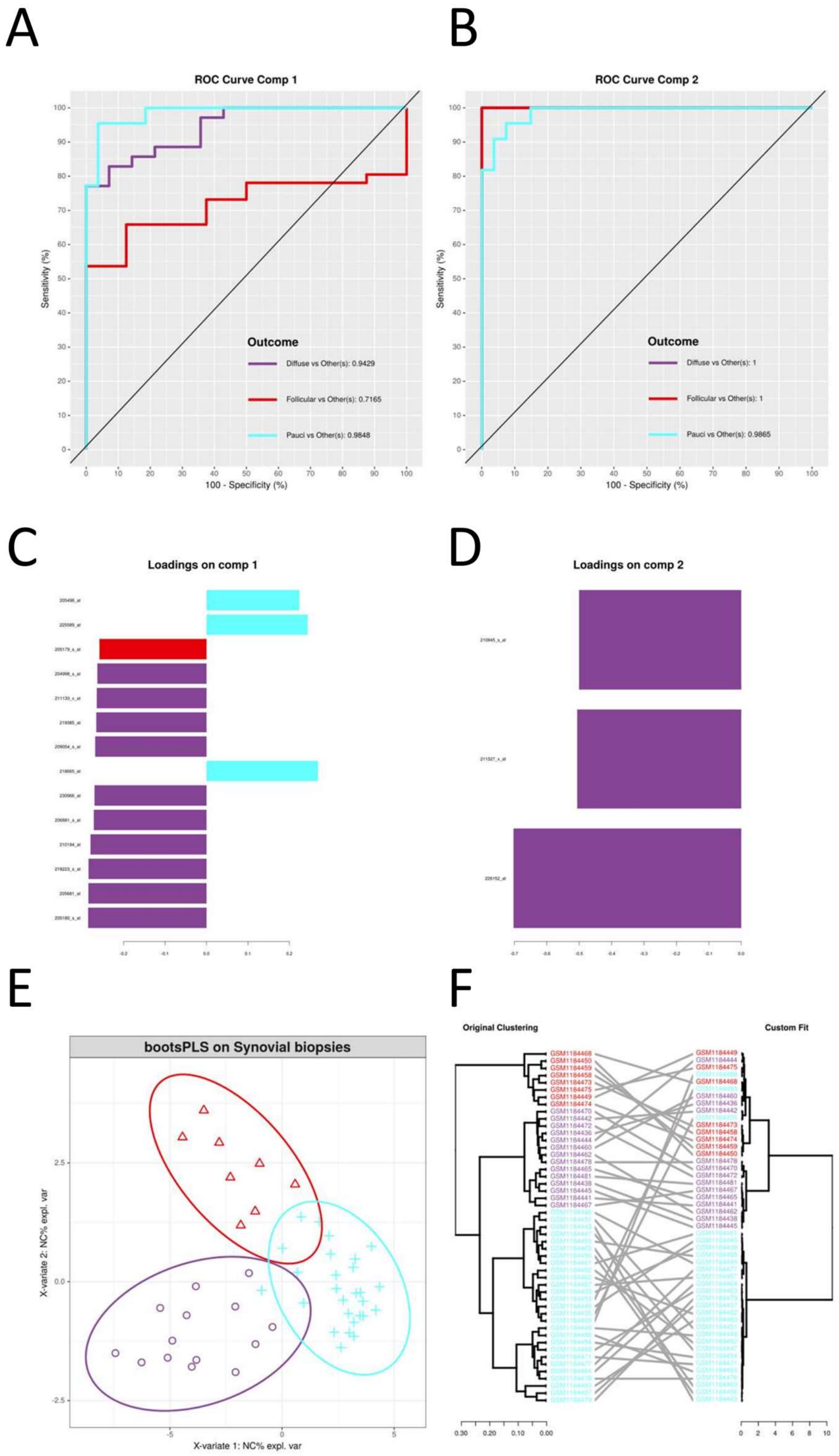


Figure 4.14: Performance of a composite signature derived from the core features of all fit models as a method to stabilise selection.

*Plotting the Receiver Operator Curves to show the sensitivity and specificity of the model utilising the first component (**A**) or second component (**B**). Performance was assessed using the Area Under the Curve, where 1 would be a perfect classification, and 0.5 would be completely random classification. The contribution of the genes to their respective components is illustrated in **C** & **D**, with the colour illustrating the pathology with the maximal mean value. **E** Plotting the variance that these probes contribute in their respective components shows a separation of the samples into their respective pathologies, ellipses reflect the 95% confidence interval for the different groups. **F** Clustering the samples has mixed ability to discriminate the pathologies when performing unsupervised clustering, with some misclassification of the Pauci samples as well as mixing the Diffuse and Follicular branches.*

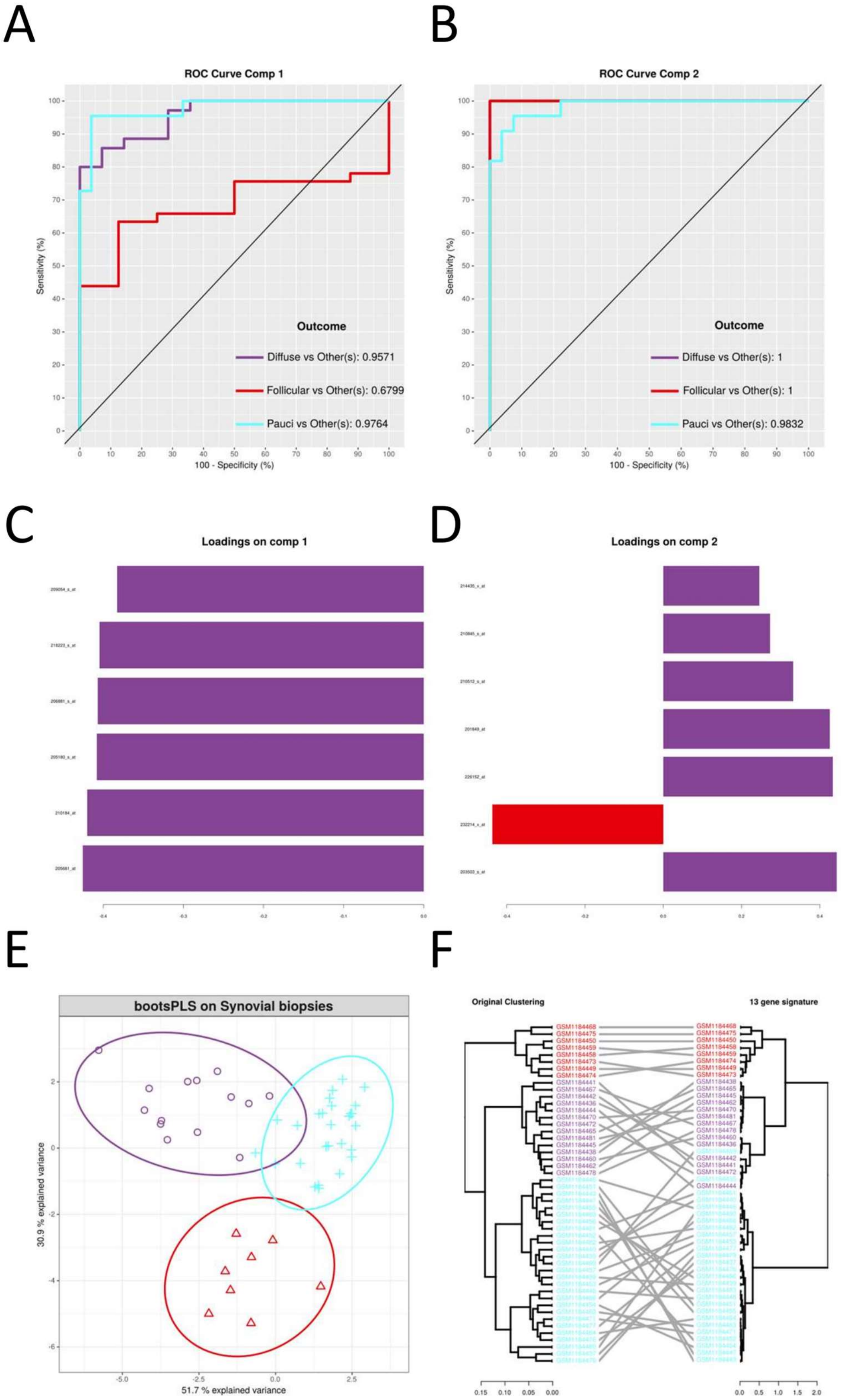


Figure 4.15: Investigating the performance of a signature identified in the pilot study into the technique.

*13 genes stratify pathologies **A, B** Receiver operator characteristics on the first and second components, and their loadings (**C, D**). **E** Plotting the variance shows good separation across the two components and good separation when performing unsupervised clustering (**F**)..*

Table 4.2: Genes utilised in the signatures identified by sPLS-DA.

Eleven of the 100 fit models were assessed as having good performance. Despite sharing the majority of probes with the good fits, fit4948 performs poorly when looking at unsupervised clustering results. These probes overlap with the initial 13 gene signature identified in the pilot study, and a 17 gene core signature common to all good fits was identified.

		Good Fit Models											“Bad”		
Probe	**Symbol**	**1556**	**2078**	**4657**	**5184**	**5237**	**5495**	**6104**	**6747**	**8395**	**8913**	**8945**	**4948**	**Core**	**13 gene**
205180_s_at	ADAM8	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
205681_at	BCL2A1	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
218223_s_at	PLEKHO1	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
210184_at	ITGAX	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
230966_at	IL4I1	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✗
205179_s_at	ADAM8	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✗
206881_s_at	LILRA3	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
209054_s_at	NSD2	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
205498_at	GHR	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✗
225589_at	SH3RF1	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✗
218665_at	FZD4	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✗
211133_x_at	LILRB3	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✗
204998_s_at	ATF5	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✗
219385_at	SLAMF8	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✗
203047_at	STK10	✓	✓	✓	✓	✓	✓	✓	✓	✗	✓	✓	✓	✗	✗
244654_at	MYO1G	✓	✓	✓	✓	✓	✓	✓	✓	✗	✓	✓	✗	✗	✗
210784_x_at	LILRB3	✓	✓	✓	✓	✓	✓	✓	✓	✗	✓	✓	✗	✗	✗
229625_at	GBP5	✗	✗	✗	✗	✓	✗	✗	✗	✗	✗	✗	✗	✗	✗
211527_x_at	VEGFA	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✗
210845_s_at	PLAUR	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
226152_at	TTC7B	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
214435_x_at	RALA	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✗	✓

202679_at	NPC1	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✗	✗
208075_s_at	CCL7	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✗	✗
210512_s_at	VEGFA	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✗	✓
201849_at	BNIP3	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✗	✓
211924_s_at	PLAUR	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✗	✗
232214_x_at	ZNF554	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✗	✓
244856_at	NA	✓	✗	✓	✗	✓	✗	✗	✓	✓	✓	✗	✗	✗	✗
55081_at	MICALL1	✗	✗	✗	✗	✓	✗	✗	✗	✓	✗	✗	✗	✗	✗
243968_x_at	FCRL1	✗	✗	✗	✗	✓	✗	✗	✗	✓	✗	✗	✗	✗	✗
38521_at	CD22	✗	✗	✗	✗	✓	✗	✗	✗	✓	✗	✗	✗	✗	✗
203719_at	ERCC1	✗	✗	✗	✗	✓	✗	✗	✗	✓	✗	✗	✗	✗	✗
203503_s_at	PEX14	✗	✗	✗	✗	✓	✗	✗	✗	✓	✗	✗	✗	✗	✓
1555425_x_at	SSH2	✗	✗	✗	✗	✓	✗	✗	✗	✓	✗	✗	✗	✗	✗
212171_x_at	VEGFA	✗	✗	✗	✗	✓	✗	✗	✗	✓	✗	✗	✗	✗	✗
214974_x_at	CXCL5	✗	✗	✗	✗	✗	✗	✗	✗	✓	✗	✗	✗	✗	✗
213927_at	MAP3K9	✗	✗	✗	✗	✗	✗	✗	✗	✓	✗	✗	✗	✗	✗
236449_at	CSTB	✗	✗	✗	✗	✗	✗	✗	✗	✓	✗	✗	✗	✗	✗
224880_at	RALA	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✓

4.6. Discussion.

It is clear that the different pathologies of RA have their own unique transcriptional profiles, which aligns with the previous investigations, as would be expected, these profiles are enriched with genes involved in pathways reflective of the biology of the pathotypes.

Given the enrichment for T- and B- cells and the formation of germinal centres in the follicular pathology, the strong associations with T & B cell receptor signalling fits well with understood biology of the pathotype. Similarly, the enhanced proliferation and angiogenesis pathways observed in the pauci samples reflects the increased pannus formation observed in the joint, highlighting that the stromal cells have a key role in disease progression. Within the diffuse pathology, we see numerous myeloid associated pathways that reflect the increased myeloid component observed in the joint. Interestingly we also see that IL-17 responses and an important component of the diffuse pathology, that were previously described as associated with the follicular pathology(88,149).

The next chapter will focus more on testing these signatures identified here, but it is important to note the novelty of this approach. Prior investigations have identified that the pathologies have their own unique molecular profiles, yet necessitated the use of the majority of the transcriptome to differentiate them(73,130). Having identified a much smaller subset of genes that discriminate between the pathologies allows for the possibility of a diagnostic test that when used in conjunction with existing knowledge of how pathologies respond to therapeutics will allow for more tailored medicine for the patients. Identifying opportunities for precision medicine has several potential positives: primarily by identifying the appropriate treatment early, it is possible to prevent irreversible joint damage from occurring; moreover, by avoiding non-efficacious treatments it minimises the negative aspects of immunomodulatory treatments and avoids potential negative outcomes that may be associated with them; and thirdly it minimises the expenditure on expensive biologic therapeutics that would be of no benefit to the patient.

One thing that is important to note with this approach, however, is that these samples represent late-stage RA, which is fully differentiated and represents patients who have been treated with unknown medication. At the start of this project, these patients represented the largest single cohort of RA patients on a modern platform, as will be explored later, newer datasets are becoming available which provide additional samples and better metadata. Moreover, if pathology becomes a focus in treatment decisions, it would seem likely that longer-term studies like R4RA will allow for better exploration of identifying signatures that predict response to therapeutics. Additionally, there is evidence for RA pathologies not being discrete and being more of a spectrum, therefore looking forwards it would be advantageous to see if it is possible to identify these “grey area” patients and see if these have different outcomes.

5. Testing signature in independent clinical datasets.

5.1. Introduction.

The advent of high throughput bioinformatic analysis of transcriptomic datasets (derived by microarray or RNA-seq) from clinical studies has opened opportunities to identify prognostic signatures that inform patient outcomes. Primarily this has been used in oncology(150) and transplantation(151) to identify signatures that define response to therapeutics, metastatic propensity, clinical features, or organ rejection. In the case of rheumatoid arthritis, high-throughput screening approaches have been extensively used to investigate synovial histopathology or the phenotypic properties of fibroblast-like synoviocytes and circulating blood leukocytes. These studies have identified several gene signatures that discriminate between rheumatoid arthritis and other forms of disease or healthy controls (152–154), as well as highlighting differentially expressed genes that give insights into the mechanisms controlling the disease(155,156). However, one of the biggest challenges is translating these discoveries into clinical practice as these signatures often fail to replicate across independently generated datasets(157).

Clinical decisions on the best course of rheumatoid arthritis therapy involve the tracking of a prescribed regime of therapeutics to determine efficacy against defined clinical outcomes and recommended NICE guidelines (16,49,50,158), outlined in Figure 5.1. This necessitates a longitudinal assessment of clinical responses to therapy for three to six months following the start of a new course of treatment (12,13), unless contraindicated by an adverse reaction. Any lack of response can lead to the progression of irreversible joint damage or uncontrolled synovitis(159). Significantly, patients may also experience a lack of therapeutic response towards multiple biological drug regimes. Here, the therapeutic strategy is geared towards maintaining the disease in a low inflammatory state and is designed to ensure the best quality of life outcome for the patient (160).

Prior studies have retrospectively identified an association between therapeutic response and histological joint pathology or immune cell phenotype(84,86,89,91,161–166). Here, ongoing clinical trials (R4RA, STRAP) with collaborators in Queen Mary University London have investigated pathology specific responses to biological therapeutics in rheumatoid arthritis (167–169). Currently, joint pathology is not assessed as part of clinical treatment decisions, partially due to the complex method of sample preparation and histological analysis. Here, early and effective diagnosis of the underlying pathology is proposed as a major decision-making tool in tailoring treatment options to enhance the likely response to biological drug therapy.

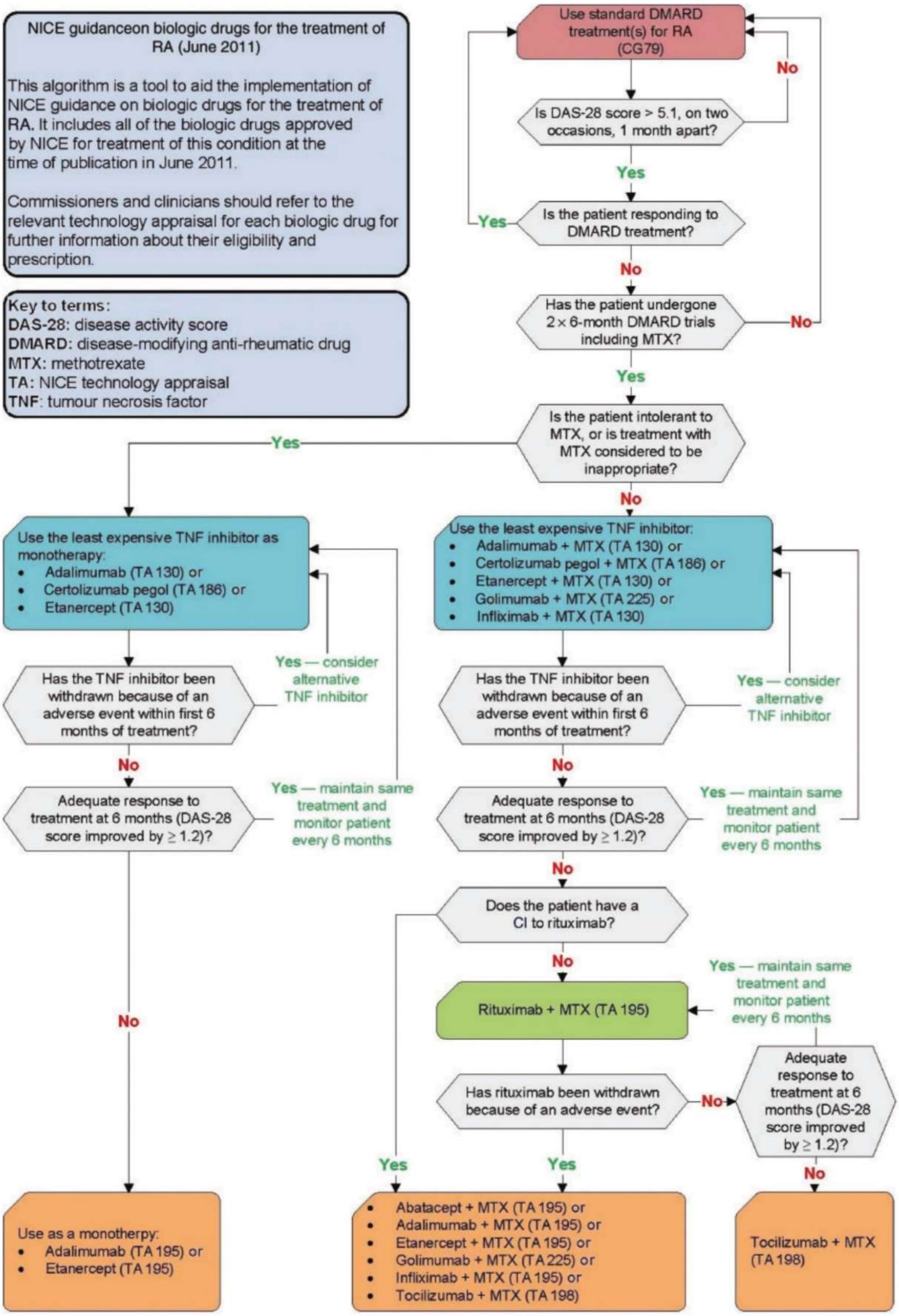

NICE guidanceon biologic drugs for the treatment of RA (June 2011)
This algorithm is a tool to aid the implementation of NICE guidance on biologic drugs for the treatment of RA. It includes all of the biologic drugs approved by NICE for treatment of this condition at the time of publication in June 2011.
Commissioners and clinicians should refer to the relevant technology appraisal for each biologic drug for further information about their eligibility and prescription.
Key to terms:
DAS-28: disease activity score
DMARD: disease-modifying anti-rheumatic drug
MTX: methotrexate
TA: NICE technology appraisal
TNF: tumour necrosis factor
Use standard DMARD treatment(s) for RA (CG79)
No
Is DAS-28 score > 5.1, on two occasions, 1 month apart?
Yes
Yes
Is the patient responding to DMARD treatment?
No
Has the patient undergone 2 × 6-month DMARD trials including MTX?
No
Yes
Is the patient intolerant to MTX, or is treatment with MTX considered to be inappropriate?
Yes
No
Use the least expensive TNF inhibitor as monotherapy:
• Adalimumab (TA 130) or
• Certolizumab pegol (TA 186) or
• Etanercept (TA 130)
Use the least expensive TNF inhibitor:
• Adalimumab + MTX (TA 130) or
• Certolizumab pegol + MTX (TA 186) or
• Etanercept + MTX (TA 130) or
• Golimumab + MTX (TA 225) or
• Infliximab + MTX (TA 130)
Has the TNF inhibitor been withdrawn because of an adverse event within first 6 months of treatment?
Yes — consider alternative TNF inhibitor
Has the TNF inhibitor been withdrawn because of an adverse event within first 6 months of treatment?
Yes — consider alternative TNF inhibitor
No
No
Adequate response to treatment at 6 months (DAS-28 score improved by ≥ 1.2)?
Yes — maintain same treatment and monitor patient every 6 months
Adequate response to treatment at 6 months (DAS-28 score improved by ≥ 1.2)?
Yes — maintain same treatment and monitor patient every 6 months
No
Does the patient have a CI to rituximab?
No
No
Rituximab + MTX (TA 195)
Yes — maintain same treatment and monitor patient every 6 months
Has rituximab been withdrawn because of an adverse event?
No
Adequate response to treatment at 6 months (DAS-28 score improved by ≥ 1.2)?
Yes
Yes
No
Use as a monotherpy:
• Adalimumab (TA 195) or
• Etanercept (TA 195)
• Abatacept + MTX (TA 195) or
• Adalimumab + MTX (TA 195) or
• Etanercept + MTX (TA 195) or
• Golimumab + MTX (TA 225) or
• Infliximab + MTX (TA 195) or
• Tocilizumab + MTX (TA 198)
Tocilizumab + MTX (TA 198)

Figure 5.1: Algorithm illustrating NICE guidance on biologic drugs for the treatment of RA.

NICE (2011) algorithm: 'rheumatoid arthritis'. www.nice.org.uk. Algorithm was accurate at the time of publication CI: contraindication. Figure from Kiely, P.D.W., et al. 2012

Histological assessments of joint biopsies provide details of cellular hyperplasia, leukocyte infiltration and the organisation of immune cells within the inflamed synovium. Whilst these approaches are labour intensive, they have opened further investigations into the transcriptomic mechanisms that drive disease heterogeneity in rheumatoid arthritis patients. Here, a prognostic gene signature that stratifies joint pathology (i.e. a disease classifier) would provide enhance treatment decisions by ensuring patients receive the most efficacious therapy for their form of the disease.

5.2. Hypothesis.

Based on the computational models described in Chapter 4, it was hypothesised that these analytical tools will aid the stratification of rheumatoid arthritis patients studied in independently evaluated clinical cohorts.

5.3. Aims.

Previous investigations have detected gene signatures that define therapeutic responses(82,144,145), but these determinations often fail to be replicated in independent datasets from alternate patient cohorts. Therefore, to test that the classifier identified in the previous chapters are not the result of overfitting, signatures derived from Chapter 4 were tested against 2 independent cohort studies that have comprehensive metadata that authentications their accuracy as validation datasets. with an overall aim to identify which gene signature displays the best performance in stratifying patients according to synovial joint pathology.

5.4. Materials and Methods.

5.4.1. Repository datasets and associated metadata.

The synovial transcriptomic datasets presented in Chapter 3 (Tables 2.4.1 & 2.4.2) were generated using a diverse selection of platforms – various microarray platforms and RNA-seq technologies. However, only two datasets contained sufficient metadata to allow comparison of transcriptomic data with joint histopathology – identified as GSE24742 and E-MTAB-6141. To allow interpretations across platforms, datasets were transformed to negate issues associated with batch effects. In this regard, all data was scaled relative to itself and not between datasets.

GSE24742 evaluated the clinical response of 12 patients to rituximab in matched samples taken before and after treatment (24 samples total). A full histological assessment of the joint biopsies derived from this study is presented as supplementary information (Tables 1a-c) within the original article publication(145). Importantly, transcriptomic datasets generated from this study were generated using the same Affymetrix HGU133plus2 microarray platform as used to produce the training data described in Chapter-4(73). Whilst the synovial pathology observed in this patient cohort was not classified into specific pathotypes, a detailed histological characterisation of each biopsy was available for this patient cohort. Using the classification system defined in Humby *et al*. 2018, pathologies were categorised using the histological criteria presented in Table 5.1. Pathologies utilised in testing classifier performance are listed in Table 5.2 and visualised in Figure 5.2.

Table 5.1: Pathotype characterisation rules as defined by Humby et al. 2019

Using a semi-quantitative evaluation of immune infiltrate. Follicular is characterised by an enrichment of B-cell and plasma cells with the presence of aggregates; Diffuse by CD3 and CD68 infiltrate low B-cell and plasma-cells and no aggregates; Pauci by low infiltrates across the board.

	Follicular	Diffuse	Pauci
B-cell (CD20)	≥2	≤1	<1
Plasma cell (CD138)	≥2	≤2	<1
T-cell (CD3)		≥1	<1
Aggregates	+	-	-
Macrophage SL (CD68)		≥2	<2

E-MTAB-6141 documents data derived from the Pathobiology of Early Arthritis Cohort (PEAC), which investigated matched blood and synovium from treatment naïve early rheumatoid arthritis patients, for testing purposes samples were restricted to synovial samples. This dataset has comprehensive metadata (see Table 5.3) and includes defined pathologies for each of the biopsy samples. Transcriptomic data from this study was generated by RNA-seq data, which necessitated some transformation of the data for modelling purposes. Here, several of the genes listed within each of the classifier signatures aligned to an alternative human sequence. These sequences originate from the over-representation of haplotype diversity in the reference genome. For example, the 2018 annotated genome reference sequence comprised 261 alternate loci that are highly enriched in polymorphic genomic regions associated with immune-related recombination events(170,171). To control for this issue, sequencing reads were mapped against the Gencode V29 Annotations(172).

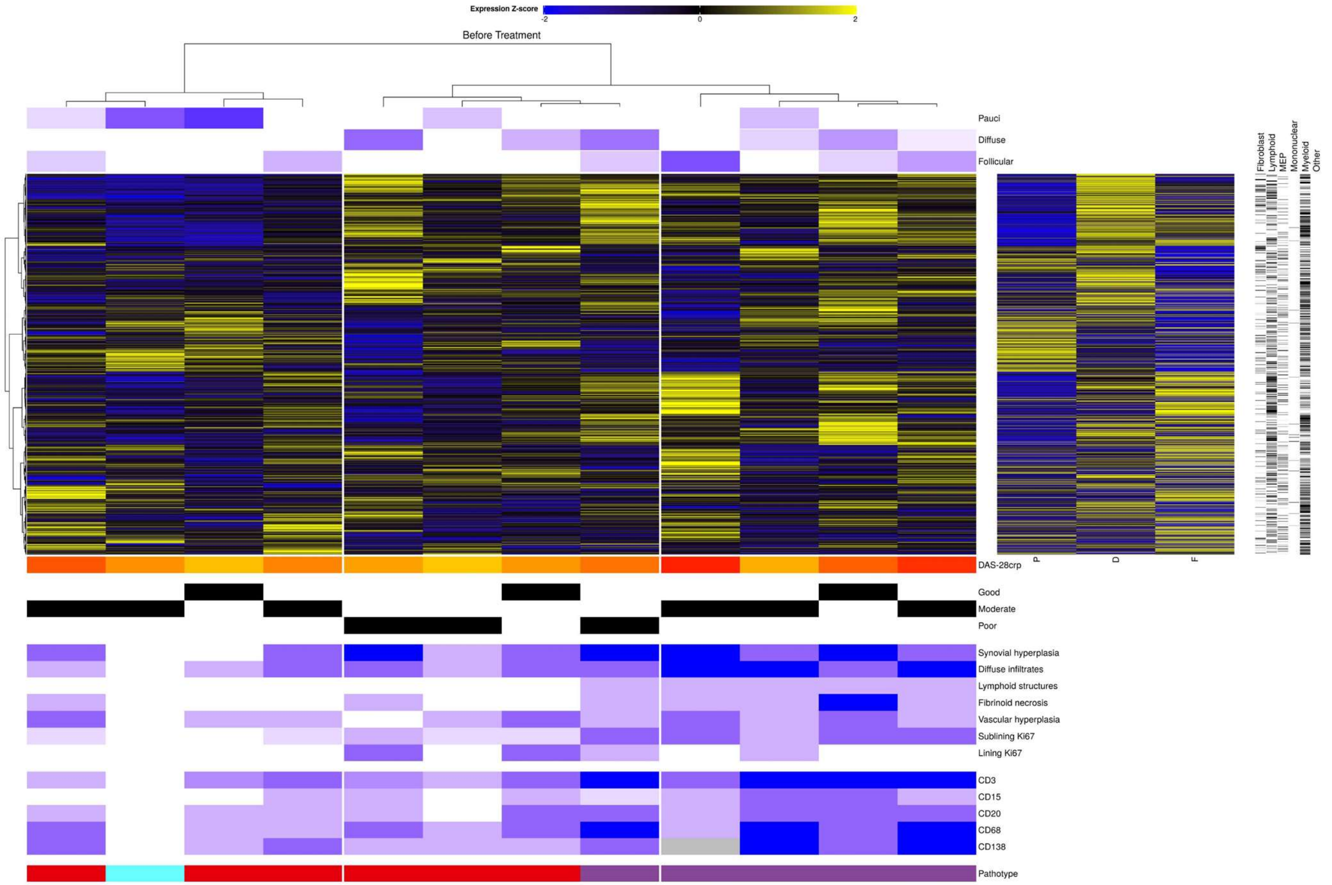


Figure 5.2: The immune profile of synovial biopsies prior to the administration of rituximab (dataset: GSE24742).

Plotting the scaled expression of these immune probes with the archetypical signature derived in chapter 4. Samples were clustered using the 13 gene signature, which reveals 3 main groups of samples. In the first cluster, the correlation between sample transcriptome and archetype shows the majority show a pauci-immune like signature, which matches with the low scores across the histological markers. Using the stricter rules of no diffuse infiltrates or lymphoid structures however reduces this to a single sample. Likewise the second cluster which strongly correlates with the diffuse archetype is almost devoid of samples containing ectopic structures. And all the samples in the third cluster show correlations with the follicular archetype, and have the presence of lymphoid structures. Pathotypes used for testing classifier performance are indicated in purple, red, and cyan – follicular, diffuse, and pauci-immune respectively

Table 5.2: Metadata of the 12 samples before rituximab treatment.

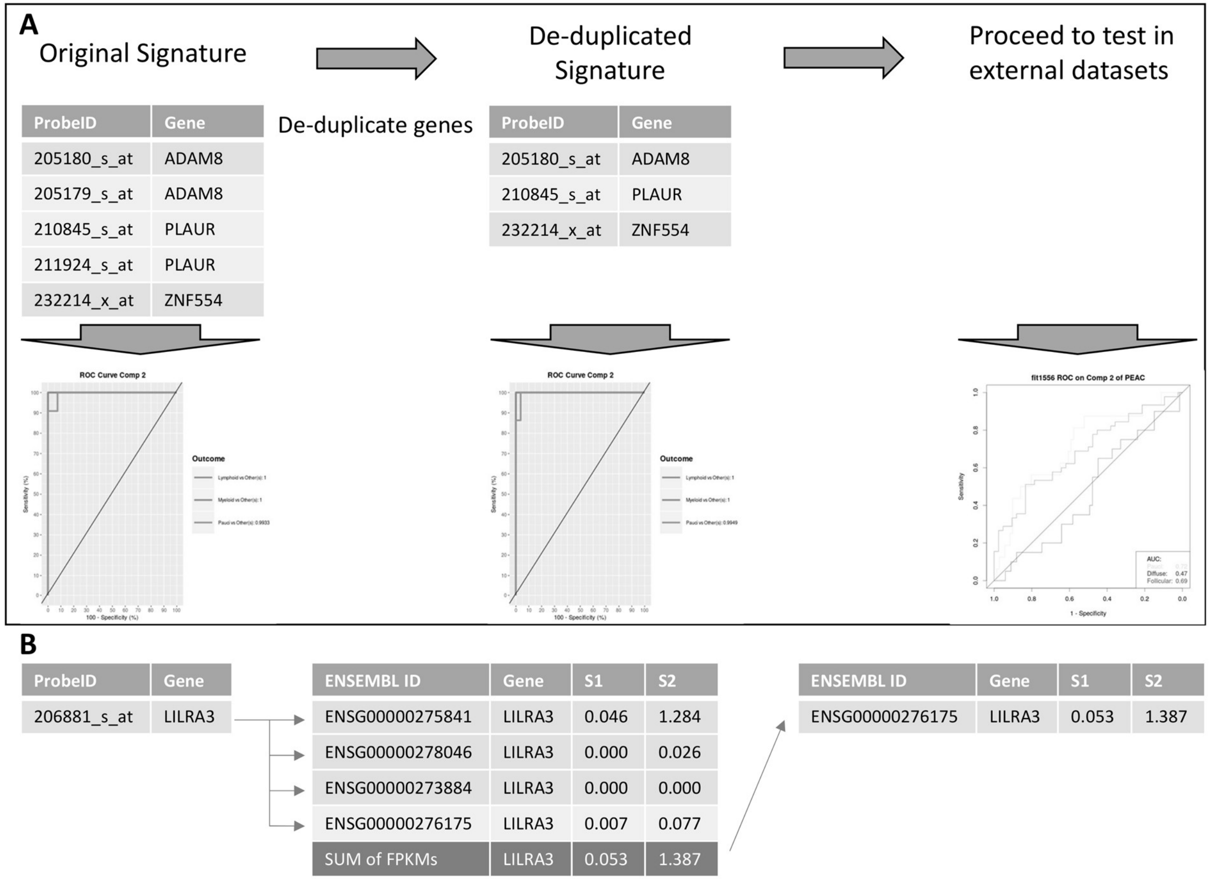
EULAR response – assessed by change in DAS28, Good, Moderate and Poor; CRP = C-Reactive Protein, VAS = Visual Analogue Scale; DAS-28 CRP = Disease Activity Score based on 28 joints and CRP levels; Histological scoring measures, Flow cytometry for CD19 (B-cells) and CD3 (T-cells); Synovial immunohistochemistry measures for specific cells – CD3 (T-Cell), CD15 (Myeloid), CD20 (B-cell), CD68 (Macrophage), CD138 (Plasma cell). Pathology was defined using rules from Humby et al 2019

Sample Name	EULAR Response	Patient	Gender	Age	Rheumatoid Factor	Anti-CCP antibodies	Swollen joints	Tender joints	CRP (mg/dl)	VAS patient	DAS-28 CRP Score	Synovial hyperplasia	Diffuse cellular infiltrates	Lymphoid structures	Fibrinoid necrosis	Vascular hyperplasia	Sub-lining Ki67 immunostaining	Lining Ki67 immunostaining	Peripheral blood CD19 (%)	Peripheral blood CD3 (%)	Synovial CD3	Synovial CD15	Synovial CD20	Synovial CD68	Synovial CD138	Pathology
GSM607508	G	1	F	51	+	+	10	9	2.2	50	5.3	2	2	0	0	2	0.5	2	17	64	2	1	2	2	1	Diffuse
GSM609031	G	2	F	64	+	+	6	5	1.6	11	4.1	0	1	0	0	1	0	0	11	69	1.5	0	1	1	1	Diffuse
GSM609033	G	3	F	68	-	-	21	24	1.6	54	6.8	3	2	1	3	2	2	0	11	72	3	2	2	2	2	Follicular
GSM609035	M	8	M	44	+	+	6	11	6.3	69	5.8	2	2	0	0	1	0.5	0	7	79	2	1	1	1	2	Diffuse
GSM609037	M	9	F	33	-	+	19	38	7	97	7.8	3	3	1	1	2	2	0	10	80	2	1	1	1	N/A	Follicular
GSM609386	M	10	M	59	-	-	13	29	2.7	80	7	2	1	0	1	2	0.5	0	3	87	1	0	1	2	2	Diffuse
GSM609388	M	11	F	77	-	-	23	26	5.1	93	7.6	2	3	1	1	1	2	0	9	84	3	1	2	3	3	Follicular
GSM609390	M	12	F	68	-	+	5	30	0.1	73	5.5	0	0	0	0	0	0	0	15	71	0	0	0	0	0	Pauci
GSM609392	M	13	F	21	+	-	4	4	1.4	71	4.6	2	3	1	1	1	1	1	14	79	3	2	2	3	3	Follicular
GSM609394	P	17	M	64	-	-	7	17	6	80	5	3	2	0	1	0	1	2	7	78	1.5	1	1	2	1	Diffuse
GSM609396	P	18	F	47	-	+	4	9	4.5	94	3.8	1	1	0	0	1	0.5	0	4	82	1	0	0	1	1	Diffuse
GSM609398	P	19	F	59	+	NA	6	6	1.2	87	6.3	3	2	1	1	1	2	1	NA	81	3	0.5	2	3	2	Follicular

Table 5.3: Sample of the metadata provided for the PEAC dataset.

In addition to pathotype, the dataset contains a wealth of clinical measures that have been explored in the original paper (Lewis, M.J. et al. 2019). Ind = Individual, to match paired blood and synovial samples; ESR = Erythrocyte Sedimentation Rate; CRP = C-Reactive Protein; CCP = cyclic citrullinated peptides or Anti-citrullinated protein antibody; RF = Rheumatoid Factor; VAS = Tender = Number of tender joints; Swollen = Number swollen joints; HAQ = Health Assessment Questionnaire disability index; DAS28 = Disease Activity Score based on 28 joints, Inf Score = Inflammatory score.

ENA_RUN	Age	Sex	Ind	Tissue	Onset	ESR	CRP	CCP	RF	VAS	TENDER	SWOLLEN	HAQ	DAS28	Inf score	Pathotype
ERR2179090	62	F	1	blood	12	106	79	179	1	48	27	26	3	8.27	4	lymphoid
ERR2179089	62	F	1	synovium	12	106	79	179	1	48	27	26	3	8.27	4	lymphoid
ERR2179092	61	F	2	blood	8	74	35	340	1	61	14	23	2	7.31	8	lymphoid
ERR2179091	61	F	2	synovium	8	74	35	340	1	61	14	23	2	7.31	8	lymphoid
ERR2179094	89	F	3	blood	12	28	9	17	1	50	4	4	NA	4.85	3	myeloid
ERR2179093	89	F	3	synovium	12	28	9	17	1	50	4	4	NA	4.85	3	myeloid
ERR2179096	71	F	4	blood	1	28	0	0	0	48	28	18	1.875	7.58	7	lymphoid
ERR2179095	71	F	4	synovium	1	28	0	0	0	48	28	18	1.875	7.58	7	lymphoid
ERR2179098	71	M	5	blood	3	95	162	600	1	92	21	24	2.875	8.47	1	fibroid



ProbeID	Gene
205180_s_at	ADAM8
205179_s_at	ADAM8
210845_s_at	PLAUR
211924_s_at	PLAUR
232214_x_at	ZNF554

ProbeID	Gene
205180_s_at	ADAM8
210845_s_at	PLAUR
232214_x_at	ZNF554

ProbeID	Gene
206881_s_at	LILRA3

ENSEMBL ID	Gene	S1	S2
ENSG00000275841	LILRA3	0.046	1.284
ENSG00000278046	LILRA3	0.000	0.026
ENSG00000273884	LILRA3	0.000	0.000
ENSG00000276175	LILRA3	0.007	0.077
SUM of FPKMs	LILRA3	0.053	1.387

ENSEMBL ID	Gene	S1	S2
ENSG00000276175	LILRA3	0.053	1.387

Figure 5.3: Workflow for testing RNA-seq data.

***A** The microarray platform utilised in the initial training has multiple probes for the same gene, whilst RNA-seq is mapped to unique genes. Therefore, probes are filtered to remove duplicate genes, keeping the most important ProbeID for each gene. As this removes the effects of these other probes, the performance of the classifier is first examined using the training data to determine if there are significant changes to the models output. Given no deleterious effects on the classifier performance, this is then taken on to testing in an external dataset. **B** Some genes exist as Human Alternative Sequences, for genes that match this, given their 100% sequence identity just alternate mapping, the FPKM values were summed and returned as a single value for further testing.*

5.4.2. Testing of classifier performance.

For predictions made using the PEAC data, signature probes corresponding to the same transcript were reduced to a single gene entry. The impact of removing these duplicated gene probes on the model was assessed by plotting Receiver Operator Curve (ROC) and comparing the Area Under the Curve (AUC) for each prediction to the training dataset.

Transcriptomic data from the respective datasets were restricted to the signature probes/genes and fed into the prediction model. Here, predictions were tested against the histologically described pathotype in a one vs all method. That is, for each pathology, samples were assigned true/false values for the pathology, and prediction confidence from the classifier was tested as case/control in ROCR(173), and this was then repeated for all pathologies, This workflow is shown in Figure 5.4 using Pauci-immune as the example pathology. To correct for discrepancies between microarray probes, gene annotation, and human alternative sequences the workflow was changed as illustrated in Figure 5.3B. Specifically, the FPKM value for genes with an alternate gene sequence was combined and returned as a sum FPKM value for each gene and entered as a single value in the classifier (Figure 5.3B). To avoid the issue of batch effects, and to make comparisons possible across platforms, all values were scaled across genes.

5.4.3. Incorporating archetypes from the microarray dataset.

GSE24742 is on the same microarray platform as the archetypes derived from GSE48780 in the previous chapters. Thus, each of the archetypes derived in Chapter-4 can be directly scaled for evaluation of the GSE24742 dataset. To explore the immune profile of PEAC dataset, the archetypical expression values of the microarray probes were converted to a compatible format to allow direct comparison with the datasets derived by microarray analysis. As with the classifier (see Section 5.4.2), genes annotated by multiple probes were reduced to a single value and renamed according to the Ensembl ID. Unlike the classifier, these probes do not have any inherent ranking of importance due to their contribution to a model. Instead, they were identified as markers of immune or stromal cells. Therefore, all probes associated with immune markers were ordered by average expression and the highest one used. In the case of multiple Ensembl IDs as a result of human alternative sequences, these ordered probes were used to fill out alternative Ensembl IDs, repeated if necessary.

Metadata

Sample	Pathology	Test.Values
ERR2179089	lymphoid	0
ERR2179091	lymphoid	0
ERR2179093	myeloid	0
ERR2179095	lymphoid	0
ERR2179097	fibroid	1
ERR2179099	fibroid	1
ERR2179101	lymphoid	0
ERR2179103	lymphoid	0
ERR2179105	lymphoid	0
ERR2179107	lymphoid	0
ERR2179109	ungraded	0
ERR2179111	fibroid	1
ERR2179113	ungraded	0
ERR2179115	myeloid	0
ERR2179117	myeloid	0
ERR2179119	fibroid	1
ERR2179121	lymphoid	0
ERR2179123	myeloid	0
ERR2179125	myeloid	0

Model predictions

Pauci	Diffuse	Follicular	Pauci
0.711159	0.215763	0.073079	0.711159
0.407506	0.203119	0.389375	0.407506
0.808627	0.268056	-0.07668	0.808627
0.914966	0.07518	0.009854	0.914966
0.838098	0.200094	-0.03819	0.838098
0.848851	0.242292	-0.09114	0.848851
0.517669	0.310911	0.17142	0.517669
0.006	0.654359	0.339641	0.006
-0.44446	0.859091	0.58537	-0.44446
0.622037	0.112003	0.265961	0.622037
0.181642	0.507681	0.310677	0.181642
0.295153	0.35415	0.350696	0.295153
1.239898	0.636804	-0.8767	1.239898
0.834007	0.036778	0.129214	0.834007
0.922907	-0.00195	0.079044	0.922907
0.581445	0.089202	0.329353	0.581445
0.275656	0.482439	0.241905	0.275656
0.718012	0.119403	0.162585	0.718012
0.615041	0.125872	0.259087	0.615041

Performance

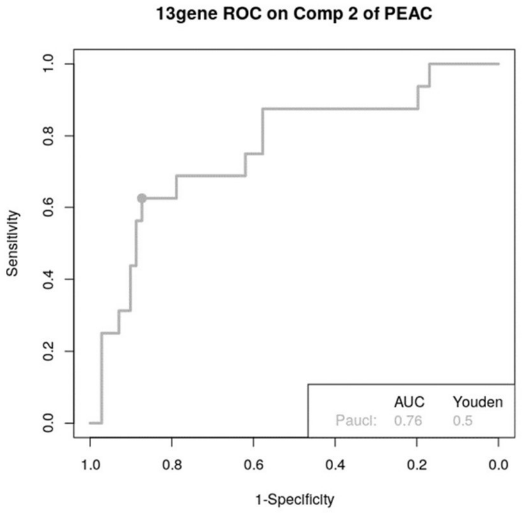


Figure 5.4: Testing rational for assessing classifier performance.

Pathologies were obtained from the metadata and tested one vs all, in this example for pauci-immune. Using the metadata, the true values for testing were reduced to true/false values, so all samples that were classified as pauci (fibroid from metadata) are assigned as one, and anything else a zero. Classifier predictions were then restricted to just the pathology being tested, and combined with the true values allow the performance to be assessed using the receiver operator characteristics, quantified using the area under the curve. This is then repeated for the other 2 pathologies.

5.5. Results.

5.5.1. The clinical datasets.

Histological assessments of the 12 biopsy samples in the GSE24742 dataset identified synovial pathologies characterised as pauci-immune (n=1; coloured cyan), follicular (n=5; coloured purple), and diffuse (n=6; coloured red) shown in Figure 5.2. The patients utilised in this cohort were all resistant to anti-TNF with a mean disease duration of 12.6 years, therefore these samples represent established disease.

The PEAC dataset provides information on 87 synovial joint biopsies characterised histologically as follicular (lymphoid; n=45), diffuse (myeloid; n=20), or pauci-immune (fibroid; n=16) synovitis. A further 6 samples had unclassified pathologies. The naming of the dataset, Pathobiology of Early Arthritis Cohort, demonstrates that these are patients with early disease, and from the patient description are all treatment naïve. The accompanying metadata provides a wealth of other clinical scores and outcome measures which have been extensively reported in several papers and used to support the investigation described here(71,130).

5.5.2. Classifier performance.

The performance of all the models generated in Chapter 4 was assessed against both independent clinical cohorts, except for the PAMR signature which fails to translate across platforms, requiring the full set microarray probes that are then assessed using a shrunken centroid methodology based on a pre-determined delta threshold, and is therefore only tested against GSE24742.

Removal of duplicate probes was assessed via AUC to ensure no drastic effects from changes to the signature, which in most cases resulted in a small increase in AUC as illustrated in Figure 5.3A

Within the GSE24742 dataset, all the models displayed a very good performance classifying pauci-immune (AUC: 0.91). The PAMR and Random Forest models have good accuracy predicting the follicular samples, but like all the models have a very poor ability to predict the diffuse pathology (AUC: 0.31~0.33). The results of all models are shown in Table 5.2.

The results of model performance within the PEAC dataset are listed in Table 5.5. However, as described above, these signatures perform well as predictors of synovial pauci-immune (AUC: 0.73~0.8) and follicular (AUC: 0.69~0.77) pathology, but as shown previously demonstrated a poor performance to predict diffuse synovitis (AUC: 0.43~0.51).

Table 5.4: Summary of the different model's performance in the GSE24742 dataset.

Model		GSE24742					
		AUC			Youden's Index		
		Follicular	Diffuse	Pauci	Follicular	Diffuse	Pauci
PAMR		0.86	0.33	0.91	0.66	0.17	0.91
Random Forest		0.73	0.31	0.91	0.31	0.17	0.91
BootsPLS	13 gene	0.63	0.31	0.91	0.43	0.17	0.91
	17 gene	0.54	0.31	0.91	0.37	0.00	0.91
	1556	0.54	0.33	0.91	0.23	0.17	0.91
	2078	0.60	0.31	0.91	0.29	0.17	0.91
	4657	0.54	0.33	0.91	0.23	0.17	0.91
	4948	0.60	0.31	0.91	0.31	0.17	0.91
	5184	0.60	0.31	0.91	0.29	0.17	0.91
	5237	0.60	0.33	0.91	0.31	0.17	0.91
	5495	0.60	0.31	0.91	0.29	0.17	0.91
	6104	0.60	0.31	0.91	0.29	0.17	0.91
	6747	0.54	0.33	0.91	0.23	0.17	0.91
	8395	0.66	0.33	0.91	0.51	0.17	0.91
	8913	0.54	0.33	0.91	0.23	0.17	0.91
	8945	0.60	0.31	0.91	0.29	0.17	0.91

Table 5.5: Model performance in the PEAC dataset.

Model		PEAC					
		AUC			Youden's Index		
		Follicular	Diffuse	Pauci	Follicular	Diffuse	Pauci
PAMR		-	-	-	-	-	-
Random Forest		0.73	0.46	0.80	0.36	0.05	0.54
BootsPLS	13 gene	0.74	0.43	0.76	0.43	0.04	0.50
	17 gene	0.77	0.51	0.77	0.45	0.17	0.46
	1556	0.69	0.48	0.75	0.32	0.14	0.45
	2078	0.69	0.48	0.75	0.32	0.14	0.45
	4657	0.69	0.48	0.75	0.32	0.14	0.45
	4948	0.67	0.48	0.73	0.32	0.15	0.42
	5184	0.69	0.48	0.75	0.32	0.14	0.45
	5237	0.78	0.48	0.78	0.50	0.15	0.50
	5495	0.69	0.48	0.75	0.32	0.14	0.45
	6104	0.69	0.48	0.75	0.32	0.14	0.45
	6747	0.69	0.48	0.75	0.32	0.14	0.45
	8395	0.76	0.48	0.75	0.45	0.14	0.46
	8913	0.69	0.48	0.75	0.32	0.14	0.45
	8945	0.69	0.48	0.75	0.32	0.14	0.45

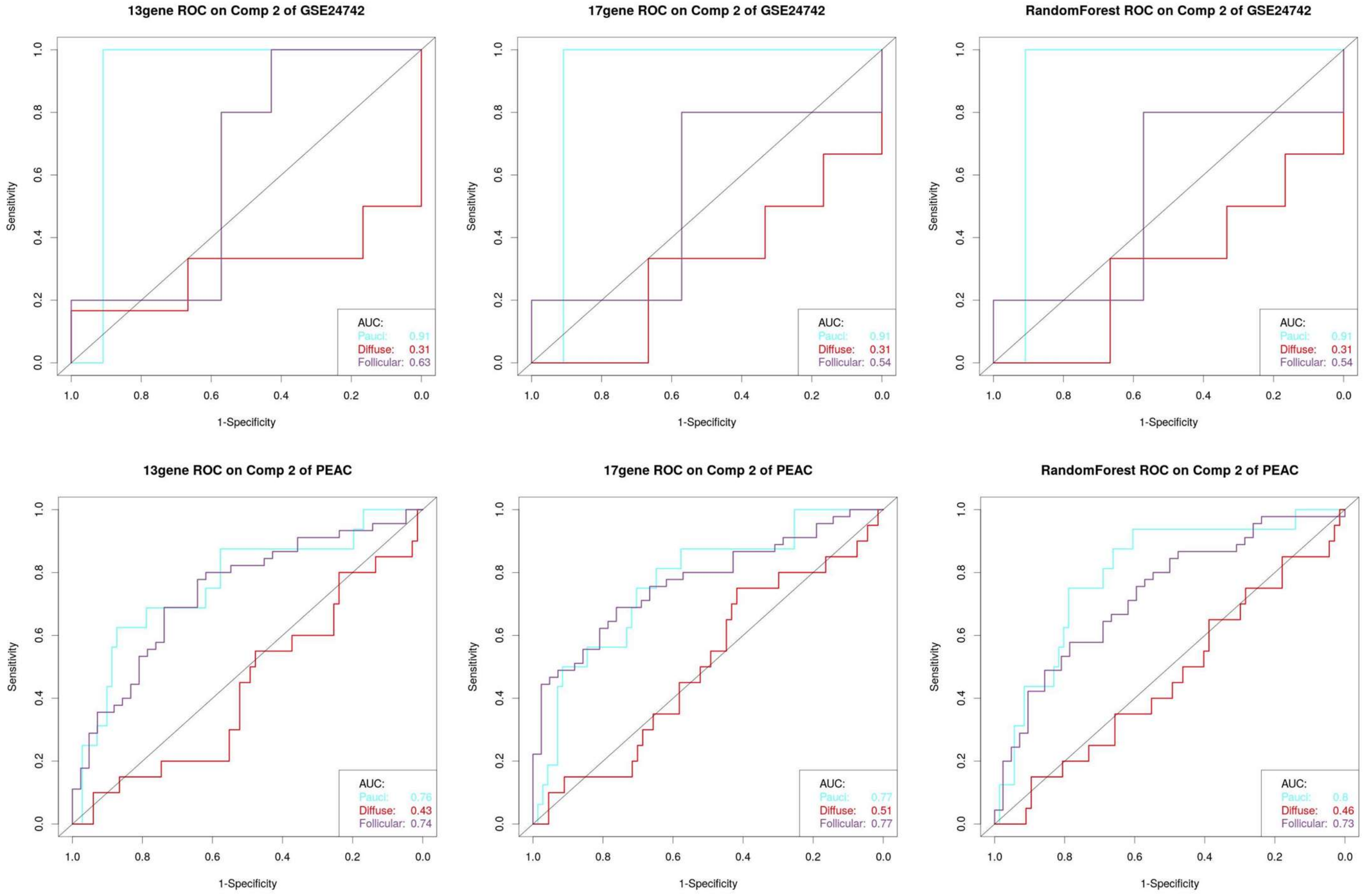


Figure 5.5: Best performing models looking in the independent clinical datasets - GSE24742 and PEAC.

All 3 models (13 gene, 17 gene, and Random Forest) show good performance for Pauci (cyan) and Follicular (purple) pathologies, but perform poorly with the Diffuse pathology (red).

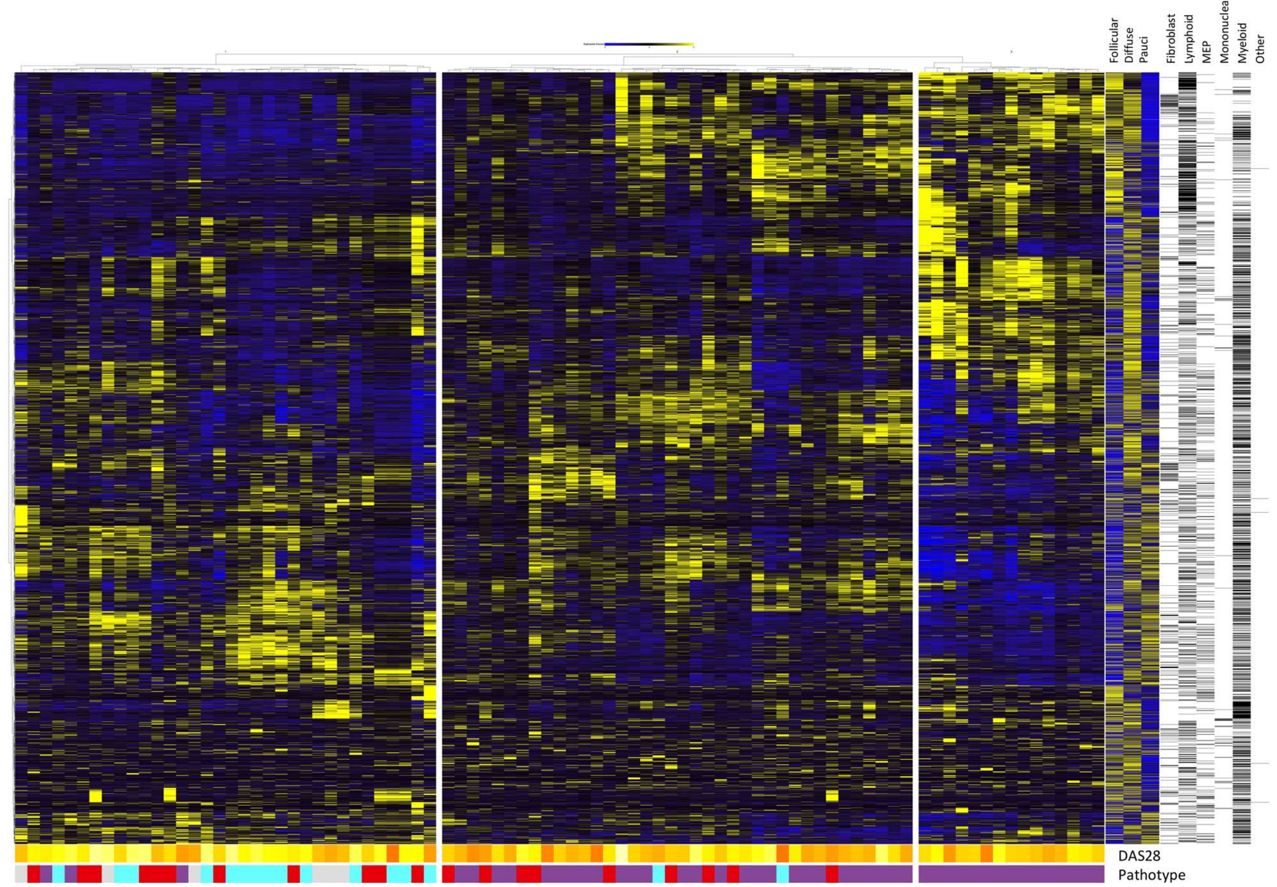


Figure 5.6: The immune profile of early arthritis.

Immune markers expressed in the PEAC dataset, compared to the archetypical expression derived in the previous chapter. Of the metadata supplied, the defined pathology and DAS28 score were plotted beneath the heatmap. Archetype expression was adapted by reducing the probes to a single gene, and mapping to the appropriate Ensembl ID. In keeping with the performance of the classifiers, there are distinct clusters that resemble the archetypes. Notably, the left cluster containing the fibroid samples behaves similarly to the to the pauci samples in the archetype, likewise the cluster on the right that contains a large proportion of the lymphoid samples shows matching behaviour to the follicular archetype. However, the cluster in the centre that contains the majority of the myeloid samples don't align to any of the archetypical behaviours.

5.5.2.1. The prediction classifier distinguishes between pauci-immune and follicular pathologies.

Comparing across both cohorts, all the models performed well at classifying the pauci immune and follicular forms of synovitis. Three models consistently performed well as classifiers of follicular and pauci-immune pathologies, both in terms of AUC and Youden's Index, which quantifies the point of maximum difference between the TPR and FPR (sensitivity and 1-specificity). The performance of these three models – Random Forest, 13 gene, and 17 gene – is plotted in Figure 5.5.

5.5.2.2. The prediction classifier is less able to stratify diffuse pathology.

All models showed poor performance when predicting the diffuse pathology, illustrated in Figure 5.5 as the red line. In the GSE24742 dataset whilst not identifying the diffuse pathology correctly, demonstrates a level of discrimination as the ROC is showing an inverted prediction behaviour, that is consistently below the 0.5 line of a truly random classifier. However, this behaviour is not seen in the PEAC dataset, where the classifier performance tracks the 0.5 line suggesting an inability to discriminate pathology.

5.5.3. Is the immune profile consistent between early-stage and to established rheumatoid arthritis?

The poor performance or inability of the classifiers to predict diffuse synovitis in these datasets, raised the question whether the diffuse pathology observed in early-stage rheumatoid arthritis showed any resemblance to the transcriptional datasets derived from established or late-stage disease and used to generate the prediction tool? To assess this, scaled datasets from biopsies with early-stage and established synovitis were related to the archetypal gene expression signature derived from the training data (Figures 5.2 and 5.6).

In Figure 5.6 there is a clear correlation between the transcriptome of the follicular archetype and the presence of ectopic lymphoid structures as shown in the top annotation of the heatmap. Likewise, the first cluster in Figure 5.6 shows samples that have very similar transcriptomes to the pauci archetype, and this is further evidenced by the histological assessments for each sample score. However, the diffuse samples show aspects of all three of the pathologies seen in the archetypical signatures, with no clear association with that of the diffuse archetype signature.

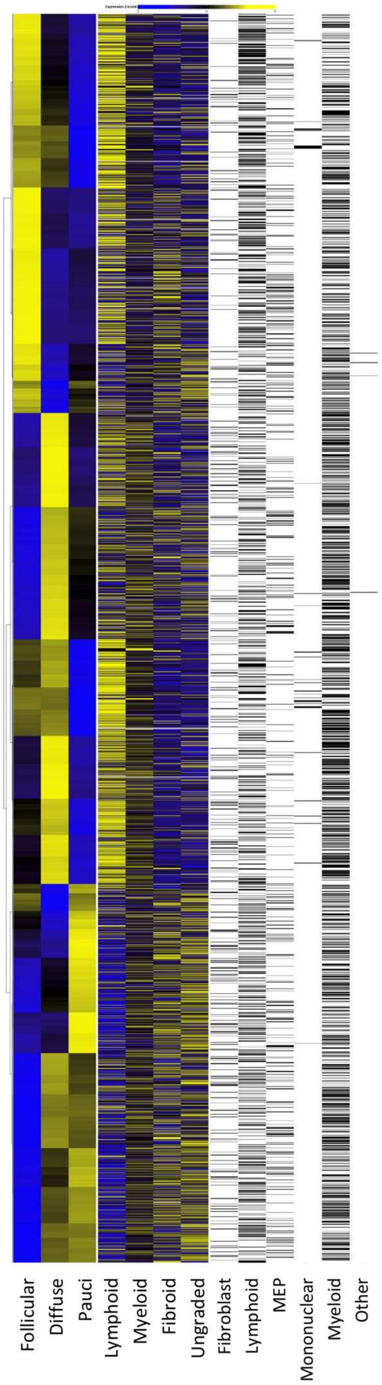


Figure 5.7: Comparing immune marker expression between archetypes derived from GSE48780 and the average expression of the pathologies as identified by histology.

The follicular immune profile is replicated in the lymphoid samples, likewise the pauci-immune profile is replicated in both the fibroid and ungraded samples. Aspects of the diffuse profile are seen in lymphoid and myeloid samples, however the myeloid sample lack a clear expression profile that matches any of the archetypes.

This is further exemplified in Figure 5.6, where the cluster on the left side of the plot that has a strong similarity to the pauci archetype, and this is evidenced with the majority of samples being identified as fibroid in the metadata – as well as a generally having a lower DAS28 score. There is also a strong correlation of follicular pathology with the cluster on the right, which is completely composed of samples identified as lymphoid in the metadata (Figure 5.6). There is also a second cluster in the centre that shows a mix of diffuse and follicular samples according to the metadata, and this is reflected with a slightly mixed pattern of upregulated markers with components of both archetypical marker genes. Looking at the samples identified as diffuse, there are no clear patterns of marker expression that matches that of the diffuse archetype. This is exemplified when looking at the average values of the PEAC samples when using the histologically defined pathologies in Figure 5.7. In addition to the defined pathologies, it can be seen that the 6 samples that were ungraded in the original data strongly resemble that of the pauci samples. And that the average myeloid expression has no strong correlation with any of the previously identified archetypes.

These established-disease patients are reflective of those used in the Dennis *et al.* training dataset, in addition to being measured on the same microarray platform. In Figure 5.2 three distinct clusters of samples match up with the immune profiles seen in the archetypes, which can be quantified by correlating the entire transcriptome between the samples and archetype. Specifically, when looking at the 2nd cluster that contains many of the diffuse samples and shows clear similarities with the diffuse archetype.

5.5.3.1. Re-evaluating the clinical definition of the pathologies.

In the original published analysis of the PEAC dataset(71,130), the authors re-evaluated the clinical definition of the synovial pathotypes to reflect the close relationship between a pauci immune synovitis and low inflammatory synovitis that reflected the enrichment of fibroblasts within the joint biopsies. Similarly, follicular-driven pathology was refined to a ‘lymphoid-myeloid’ pathotype to account for the role of myeloid cells in addition to CD20 B-cell rich lymphoid aggregates. This allowed the definition of diffuse synovitis to be characterised by the involvement of myeloid cells in the absence of a significant, organised synovial B-cell infiltrates. To consider this extra layer of sub-classification, the expression of the immune and stromal cell markers were compared between the early-stage samples and the archetypical profiles of established disease.

- Therefore, in Figure 5.8 isolating the B-cell markers shows a clear enrichment and characteristic follicular pathology for the right-hand cluster comprised of entirely follicular samples, with the central cluster showing some samples with the strong B-cell marker expression and others without. And the left-hand cluster comprising primarily of pauci samples lack this expression.
- Likewise, exploring the fibroblast markers in Figure 5.9 shows the distinct expression pattern of the pauci-immune archetype in the fibroid samples in the left-hand cluster.

- Restricting to myeloid markers in Figure 5.10 fails to recapitulate the diffuse pathotype seen in the late stage archetypical patients. Furthermore, these myeloid markers show differing components within and between clusters, but no consistent pattern associated with exclusively myeloid or diffuse patients.

In summary, both the pauci immune and follicular pathologies reflect the pattern of immune and stromal cell markers seen in the archetypical late-stage disease, but the diffuse pathology shows mixed behaviour relative to late-stage disease. Therefore, whilst there are similarities, and histological data shows that joint pathology is relatively stable between early and late disease(174,175), early diffuse synovial transcriptome has differences that will need to be accounted for to have an effective clinical tool.

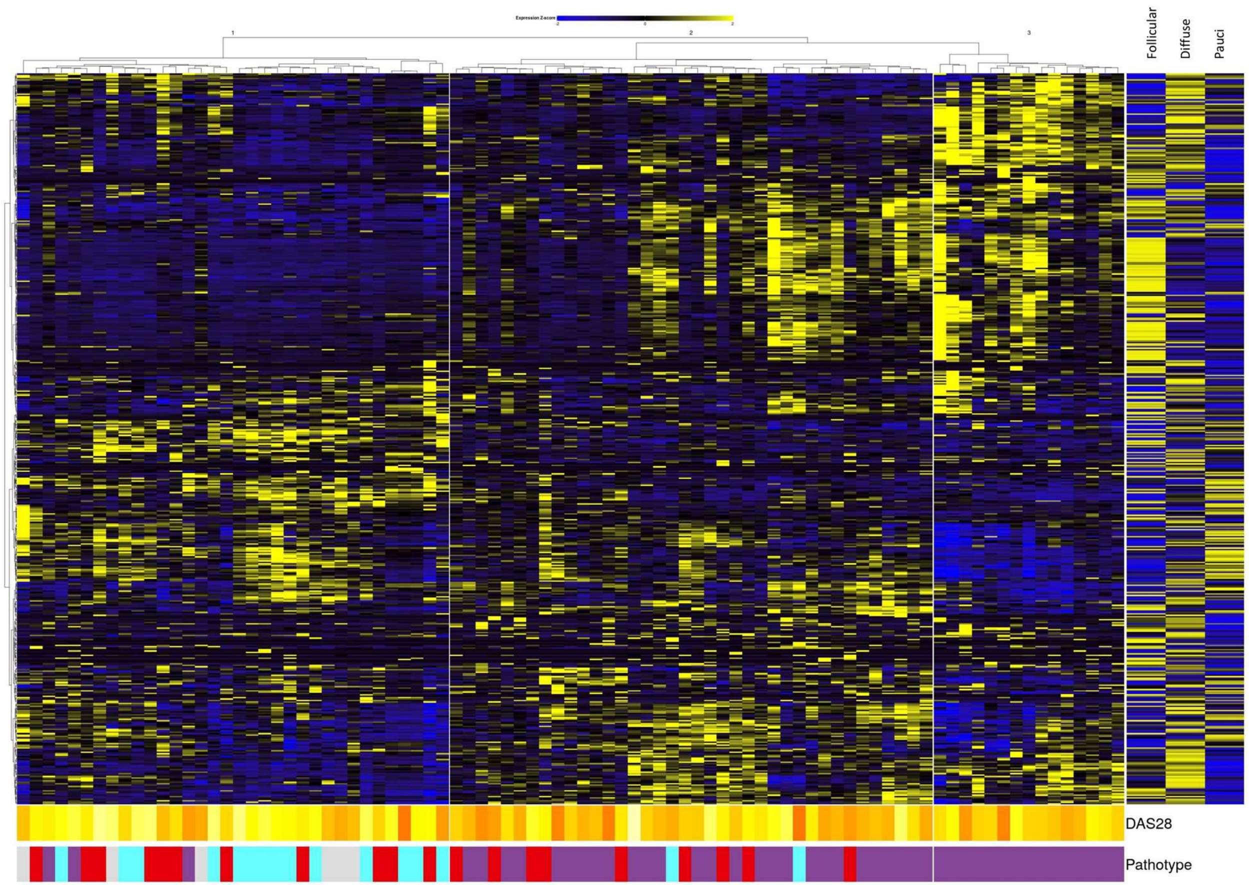


Figure 5.8: : Scaled expression of B-cell markers in the PEAC dataset relative to archetype.

Given the definition of the diffuse pathotype as being poor in CD20+ B-cells, all B-cell markers were isolated and plotted to show the strong enrichment within the Lymphoid-myeloid or follicular pathology.

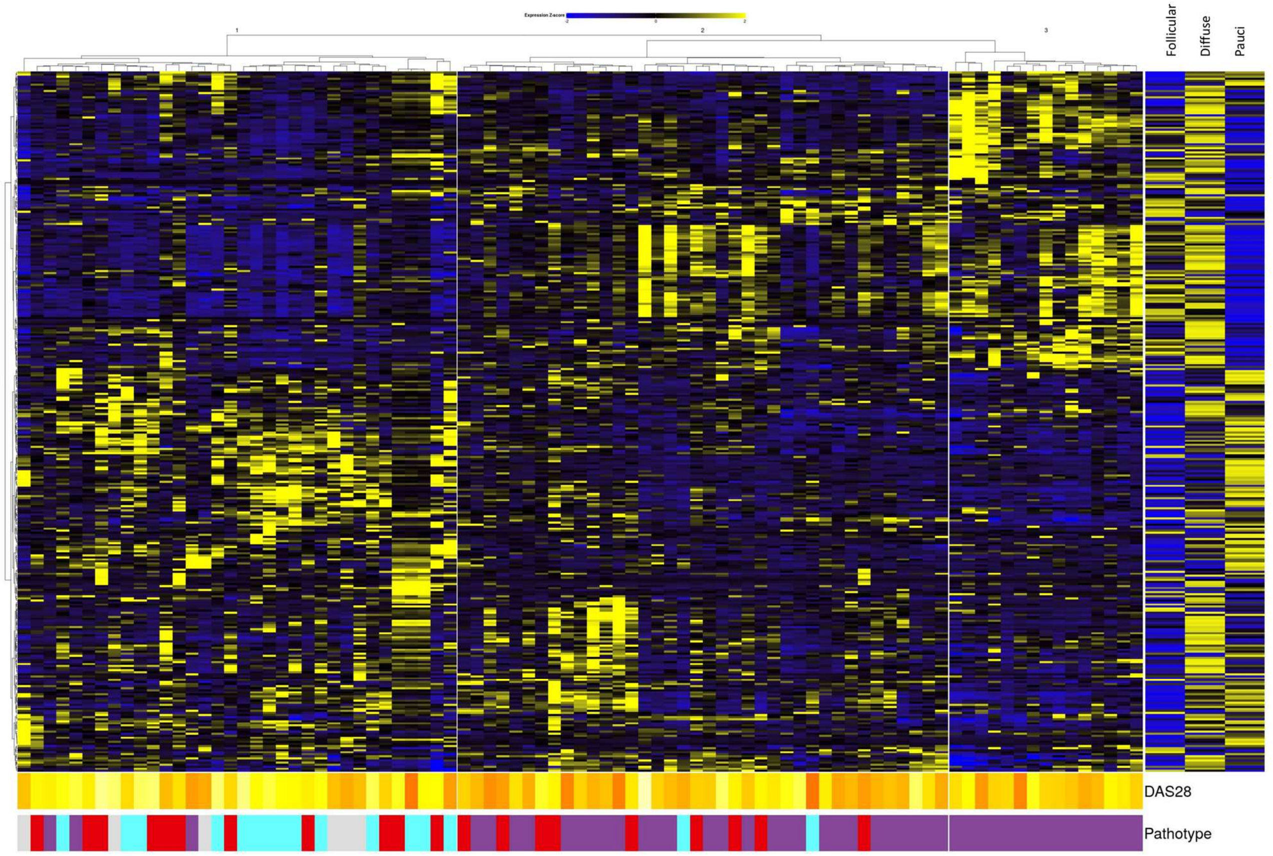


Figure 5.9 : Scaled expression of Fibroblast markers in the PEAC dataset relative to archetype.

The left cluster mainly comprising fibroid samples reflects the fibroblast signature seen in established

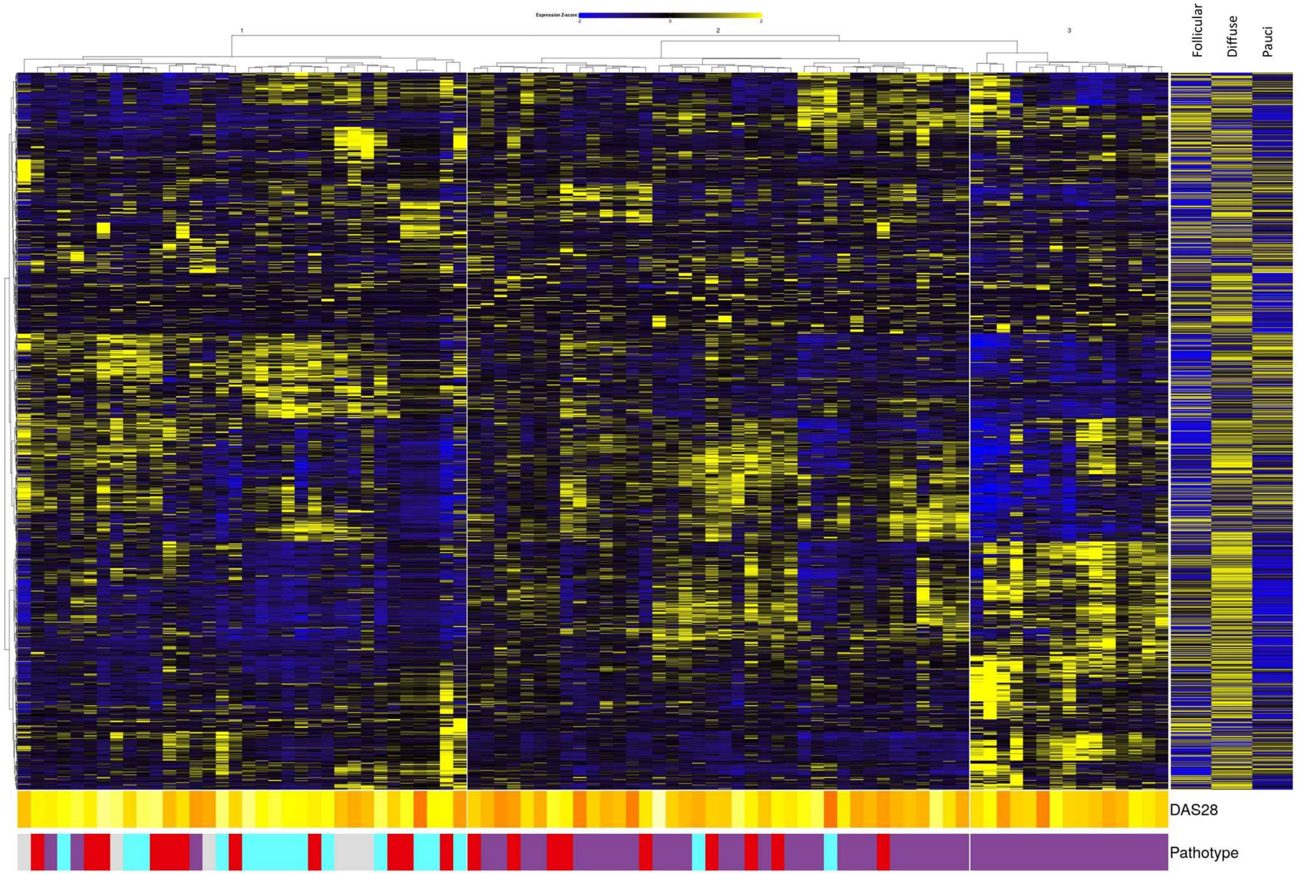


Figure 5.10: Scaled expression of Myeloid markers in the PEAC dataset relative to archetype.

Early stage RA fails to recapitulate the expression of myeloid markers seen in the established disease.

5.6. Discussion

Given that synovial pathology affects the response to biological therapies, and that early intervention is key to remission, the ability to inform therapeutic decisions with easily obtained pathology data will be key to effective precision medicine. Previous studies have demonstrated signatures that discriminate between rheumatoid arthritis and osteoarthritis or other diseases, but these often fail to be replicated in other independent studies(74,157). Here, clinical studies describing the efficacy or mode-of-action of biological therapies tend to ignore the heterogeneity of synovitis seen in rheumatoid arthritis patients and details of the joint pathology are often not included as part of the clinical assessment (144,145). This may explain why some biological therapies fail to meet their clinical endpoints. One example is secukinumab. This anti-IL-17 blocking monoclonal antibody displayed poor efficacy in rheumatoid arthritis (176–178). However, as greater insights into the potential role of IL-17 in ectopic lymphoneogenesis emerge, the effects of this biological drug on patients with lymphoid-driven pathology may have been lost due to lack of stratification. Here, a greater understanding of the biological processes driving synovial pathology combined with clinical outcome measures in routine practice is helping to formulate new clinical trials. For example, patients with lymphoid-rich synovitis typically show an inadequate response to anti-TNF inhibitors (80). This observation has led to clinical trials (e.g., R4RA and STRAP) where patients stratified according to synovial histopathology are prescribed tocilizumab or rituximab targeting either Jak-STAT signalling or depletion the involvement of B-cells in lymphoid-rich pathology (167–169). In this Chapter, experiments tested the utility of a transcriptomic classifier tool as a potential diagnostic tool that may support these therapeutic strategies.

Several of the gene signatures displayed good performance in the 2 independent datasets tested as a classifier of pauci and follicular disease, in particular the Random Forest, 13 gene, and 17 gene signatures. Unfortunately, this performance was not translated to the stratification of diffuse pathology. This inability to stratify patients with diffuse synovitis undoubtedly affects the performance of the classifier's ability to discriminate pauci-immune or follicular from that of myeloid-rich disease. As demonstrated in Figure 5.3A removal of additional probes from duplicated genes can improve the AUC, therefore additional probes involved in discriminating diffuse in the original dataset may affect the accuracy of classification. In GSE24742, the classifier exhibited an inverse predictive behaviour, so whilst incorrectly defining the diffuse pathology it was doing so more than would be expected by random. Given this behaviour, it might be tempting to correct the inversion by subtracting the AUC from 1, however, this is not replicated in PEAC dataset and therefore is not reliable to determine diffuse pathology. Furthermore, the main causes for such an inversion are not applicable: from labelling errors when creating the model (i.e. reversing the case/control definitions) or imbalances in the training/test data(179,180). However, the models were generated as a multi-class stratification and are not assigned a true/false label that could be accidentally reversed. Likewise, there are differences between the proportion of pathologies seen in training and test data sets (Train: 8 Follicular, 14 Diffuse, and 27 Pauci; and 5F,6D,1P respectively for GSE24742, and 45F,20D,16P, with addition 6 ungraded for PEAC), however, the models prove effective with the other pathologies

Here, an examination of the immune cell involvement in each of the pathologies demonstrated that this was due to an underrepresentation of myeloid-specific gene signatures within the datasets from patients with early forms of synovitis. As a consequence, myeloid-specific gene signatures were more prominent in synovial biopsies from patients with more established disease (see Figure 5.10). The data further suggests that the tracking of myeloid-specific signatures in synovial biopsies may not work as a robust diagnostic of synovial pathology, particularly during the early stage of the disease. Whilst myeloid-cells are integral to the pathology of diffuse synovitis, they are also associated with follicular disease and transcriptomic analysis of the myeloid compartment within the datasets tested showed a close association between these two forms of pathology. Moreover, resident synovial mononuclear cells will also actively contribute to all forms of synovitis including the pauci immune pathology. Such observations may explain why diffuse synovitis was difficult to predict using the classifier tool and equally suggests a closer inspection of the innate immune system is warranted to understand how myeloid cells contribute to this development of synovitis. In this regard, diffuse synovitis has recently been sub-classified to identify both fibroblast-myeloid and myeloid-lymphoid forms of synovitis(130). Thus, the form of pathology may display various ‘shades’ of pathology suggesting the need for a more detailed analysis of this form of disease.

Given the inherent differences between the transcriptional profile of the early and late-stage disease, further work will be required to evolve prediction tools that differentiate early-stage gene signatures that inform the future course of synovitis and the development of alternate synovial pathotypes. This will be explored in greater detail in Chapter-6.

In summary, the initial approach adopted here has demonstrated that effective classifiers can be utilised on datasets derived from various analytical platforms. By scaling the data this has eliminated problems arising from batch effects between the alternate datasets, and data derived using different analytical platforms. The approach may also have utility in linking analyses between mouse and human studies allowing a greater mechanistic understanding of pathways driving disease heterogeneity in rheumatoid arthritis.

6. Testing of the disease classifier in early-stage rheumatoid arthritis.

6.1. Introduction.

Clinical outcomes associated with the management of rheumatoid arthritis show that early intervention offers the best opportunity to control disease activity(49,181). Here, successful therapy with the correct drug strategy often leads to improvements in the patient quality of life and the potential drug-free remission of disease(48). Since the introduction of biological drugs, the classification of what constitutes early disease has significantly changed and may constitute as little as 6 weeks following clinical diagnosis(182). However, the definition of this 'window of opportunity' for therapy remains highly subjective and is dependent on the rate of disease progression seen in a patient or patient group. Diagnostic prediction tools must, therefore, recognise both the signs of disease progression and the pathways responsible for driving the pathology. Incorporating the bioinformatic approaches developed in this thesis, experiments described in this Chapter evaluated their utility as classifiers of early pathology.

The importance of early intervention in clinical practice is perhaps best reflected by studies on rheumatoid arthritis patients on standard DMARD therapy(183). Here, the likelihood of DMARD-free remission is reduced by delays in clinical diagnosis or referral to a rheumatologist(183)(184,185). Moreover, the prescriptive treat-to-target guidelines adopted by NICE further compromises the decision-making process and necessitates waiting to determine treatment efficacy before taking the next steps. In a move towards a more precision medicine approach as outlined by the methods described in the previous Chapters, an improved prediction of pathology at an early stage of the disease would provide valuable insights into the best course of therapy for a particular patient or patient demographic.

Previous investigations suggest that early and established disease have very similar histopathology and cytokine profiles(186–189). Evidence of this stability was illustrated through my analysis of follicular and pauci pathologies, which showed a strong correlation between early and established disease (discussed in Chapter-5). However, the classification of patients with diffuse synovitis remained more challenging due to the varying composition of immune and stromal cell involvement in this form of pathology. To potentially improve the prediction of early diffuse synovitis, experiments outlined in this Chapter sought to generate a more bespoke classifier of early disease.

6.2. Hypothesis.

Experiments outlined here will test the hypothesis that a bioinformatic classifier generated from transcriptomic data derived from early arthritis patients will improve the classification of synovial pathology. To address this hypothesis, studies used the datasets obtained through the PEAC study of early synovitis..

6.3. Aims.

As demonstrated in the previous chapter, the transcriptome of follicular and pauci pathologies are stable across both early and established disease. However, the classification of diffuse synovial pathology is more complex and variations in the cellular composition of the disease process makes the analysis more challenging (see Chapters 4 & 5). Experiments were, therefore, conducted to generate a bespoke classifier of early disease using transcriptomic datasets obtained from patients enrolled to early arthritis clinics. Here, the aim was to improve the stratification of synovial samples presenting with an early form of diffuse pathology.

6.4. Materials and Methods.

6.4.1. Repository datasets and associated metadata.

Containing a wealth of clinical outcome measures, E-MTAB-6141 (Pathobiology of Early Arthritis Cohort; PEAC), was used to generate a classifier of early-stage synovitis (130).

RNA-sequencing data available through the GSE89408 dataset provides synovial transcriptomic data from healthy, arthralgia, osteoarthritis, undifferentiated arthritis, early rheumatoid arthritis, and established arthritis. The supplementary metadata associated with this dataset offers additional information on the gender, age, and the detection of anti-citrullinated protein antibodies (ACPA) in a subset of the patient samples. A summary of the baseline characteristics of this dataset is provided in Table 6.1.

Table 6.1: Baseline patient characteristics of GSE89408.

Disease	n	Female (%)	Age (years)	ACPA Positivity (%)
Arthralgia	10	80	52.5 (33-66)	100
Healthy	28	50	35 (13-73)	NA
Osteoarthritis	22	59	49 (17-77)	NA
Rheumatoid arthritis (early)	57	58	56 (25-93)	68
Rheumatoid arthritis (established)	95	77	54 (24-85)	30
Undifferentiated arthritis	6	83	Unknown	100

6.4.2. Generating an early-stage predictive signature.

Gene signatures were explored using Sparse Partial Least Squares Discriminatory Analysis (sPLS-DA). To ensure that the bootstrapping method utilised previously in Chapter 4 worked with the datasets used here, several filtering steps were required to prepare the data. The details of data preparation are outlined in Chapter 3..2.2. Briefly, genes were filtered to remove values corresponding to pseudoautosomal region mappings and near-zero variance. To determine the optimal number of genes per component, analysis employed 400 rounds of bootstrapping.

6.4.3. Stratifying early and established disease.

As the metadata associated with the GSE89408 dataset provided no information on the precise synovial histopathology seen in each tissue extract, samples were stratified using transcriptomic data derived from immune and stromal cells (as described in Chapter 4) and archetypes derived from the PEAC cohort where transcriptomic data could be assigned to specific forms of histopathology. Through this analysis, a subset of samples with clearly defined parameters of disease was selected from the GSE89408 cohort for further testing of classifier performance.

6.4.4. Testing classifier performance.

The performance of classifiers obtained from early and established forms of synovitis was assessed for sensitivity and specificity using the prediction models described in Chapter 4 and the new signatures derived from the PEAC dataset.

6.5. Results.

6.5.1. An early-stage signature.

As discussed in Chapter 4, the models of best fit employed to test the accuracy of the prediction tools showed some degree of intervariability. As a consequence, 5 best fit models were generated. The performance of these models are summarised in Table 6.2. As the purpose of this analysis was to improve the stratification of diffuse patients, the best predictive model for this pathology is fit123, which achieves an AUC of 0.892 (Figure 6.1B). This model also has the benefit of having the smallest number of genes (how many?), which is ideal in the context of a diagnostic tool.

Table 6.2: Fit models derived from 400 iterations of bootstrapping on the PEAC dataset.

Fit represents a seed set to ensure reproducibility (see Chapter 3.1.6)

Fit	# genes (1st + 2nd component)	AUC		
		Follicular	Diffuse	Pauci
22	323	0.93	0.86	0.94
42	218	0.91	0.82	0.93
123	34	0.93	0.89	0.95
1234	102	0.93	0.87	0.95
2020	218	0.91	0.82	0.93

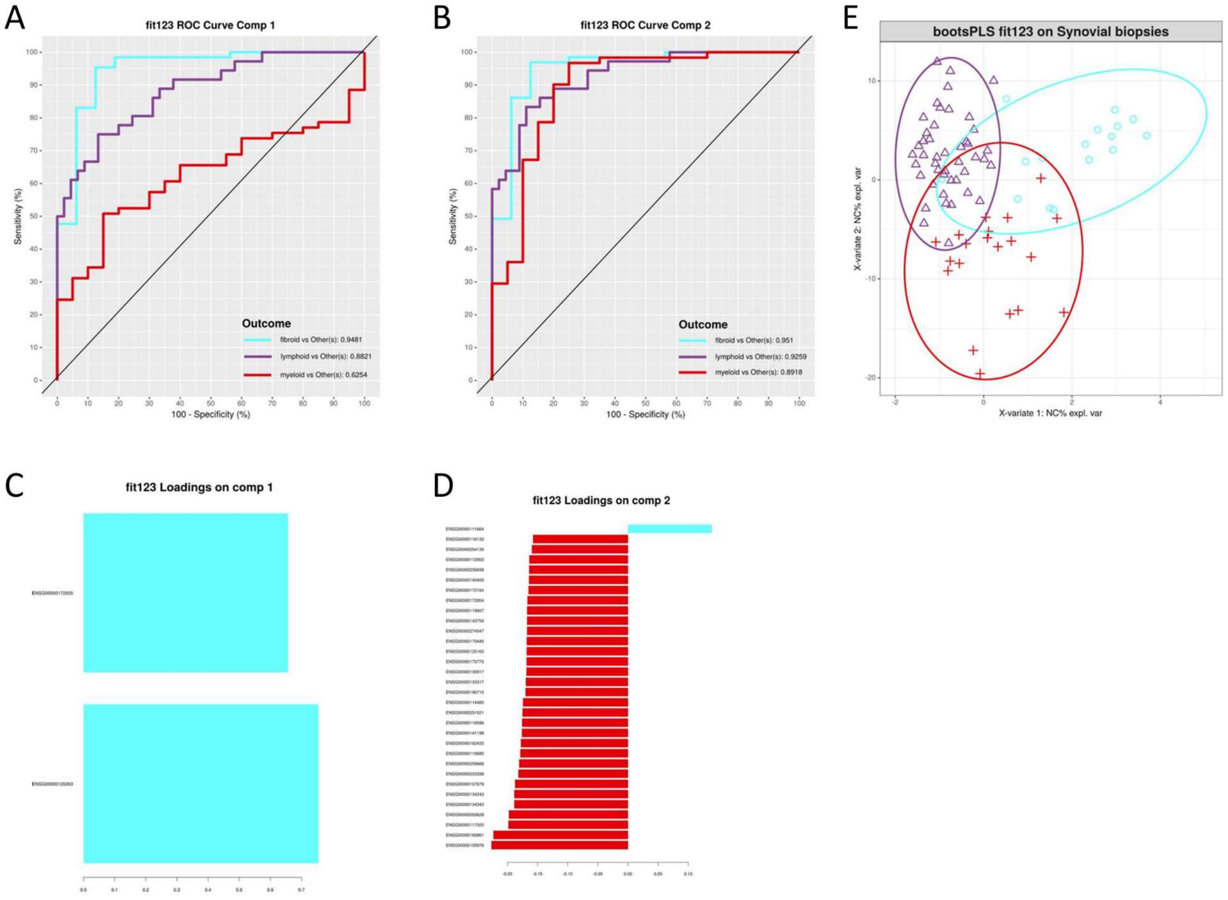


Figure 6.1: Testing performance of classifier derived from early arthritis.

*Fit123 results in a 34 gene classifier that stratifies pathologies. **A**, **B** Receiver operator characteristics on the first and second components, and their loadings (**C**, **D**). **E** Plotting the variance shows good separation across the two components*

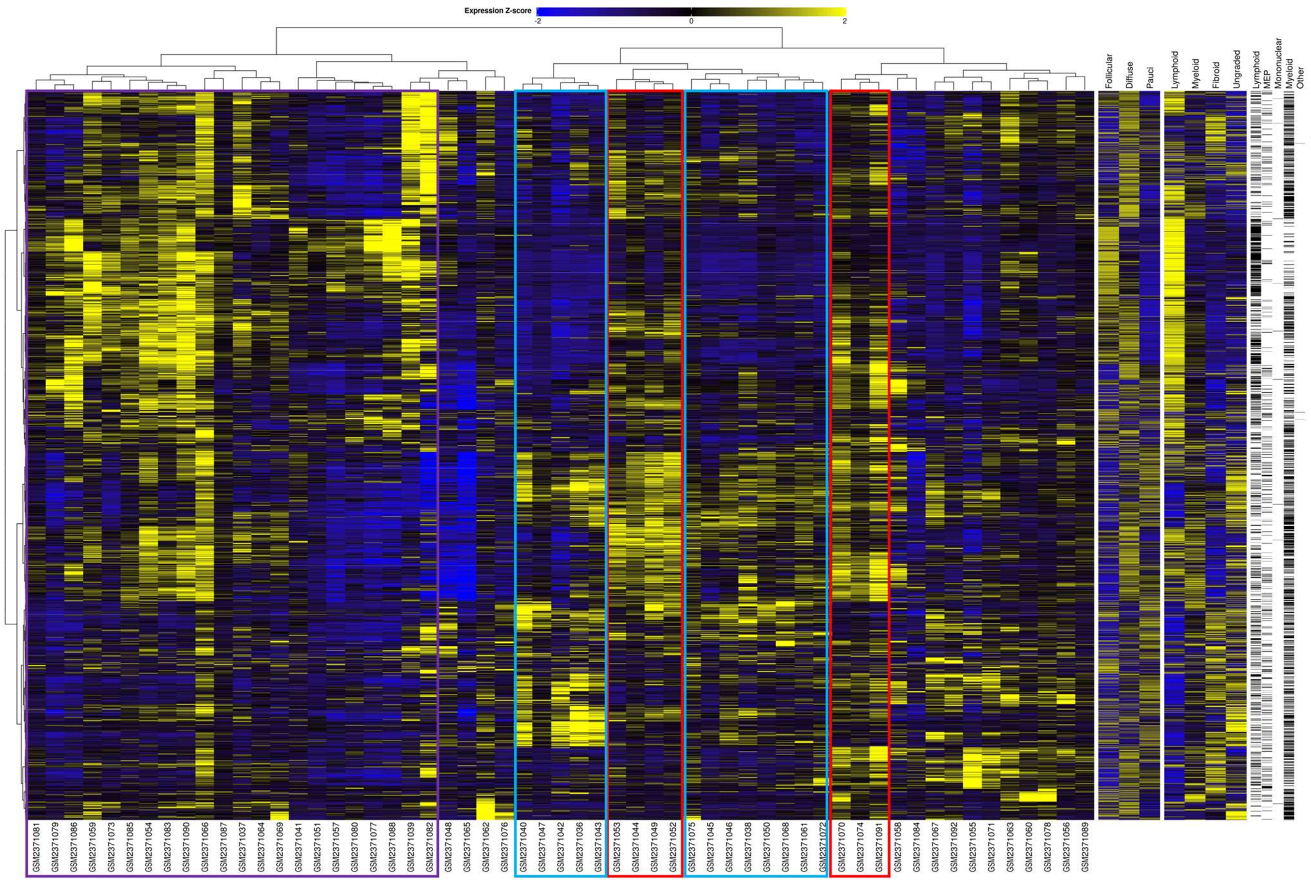


Figure 6.2: Immune and stromal markers in early Arthritis.

Patients with early rheumatoid arthritis from GSE89408 are contrasted with the archetypes derived from GSE48780 (Follicular, Diffuse, Pauci) and E-MTAB-6141 (Lymphoid, Myeloid, Fibroid, Ungraded). This dataset lacks metadata to define pathology therefore samples were characterised by similarity to the archetypes. Samples that are representative of these pathologies are denoted by the coloured box (Follicular – purple; Diffuse – red; Pauci – cyan) which was then utilised in further analysis.

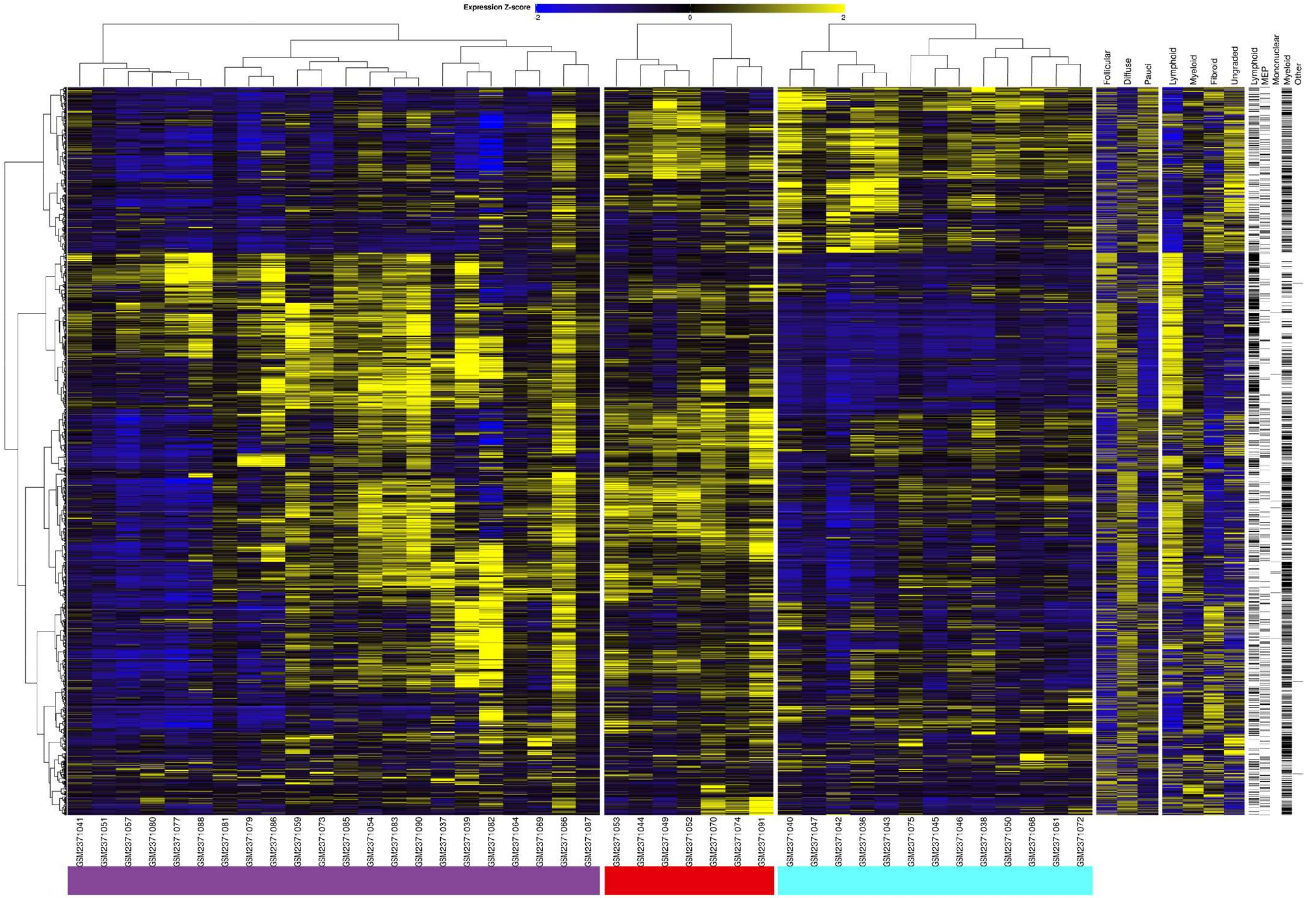


Figure 6.3: Immune and stromal markers in early arthritis with defined pathologies.

The defined samples that show clear stratification of early-stage rheumatoid arthritis patients (GSE89408) that match the immune profile of the archetypes. Resulting in 21 follicular, 7 diffuse, and 13 pauci-immune samples that can be tested going forwards.

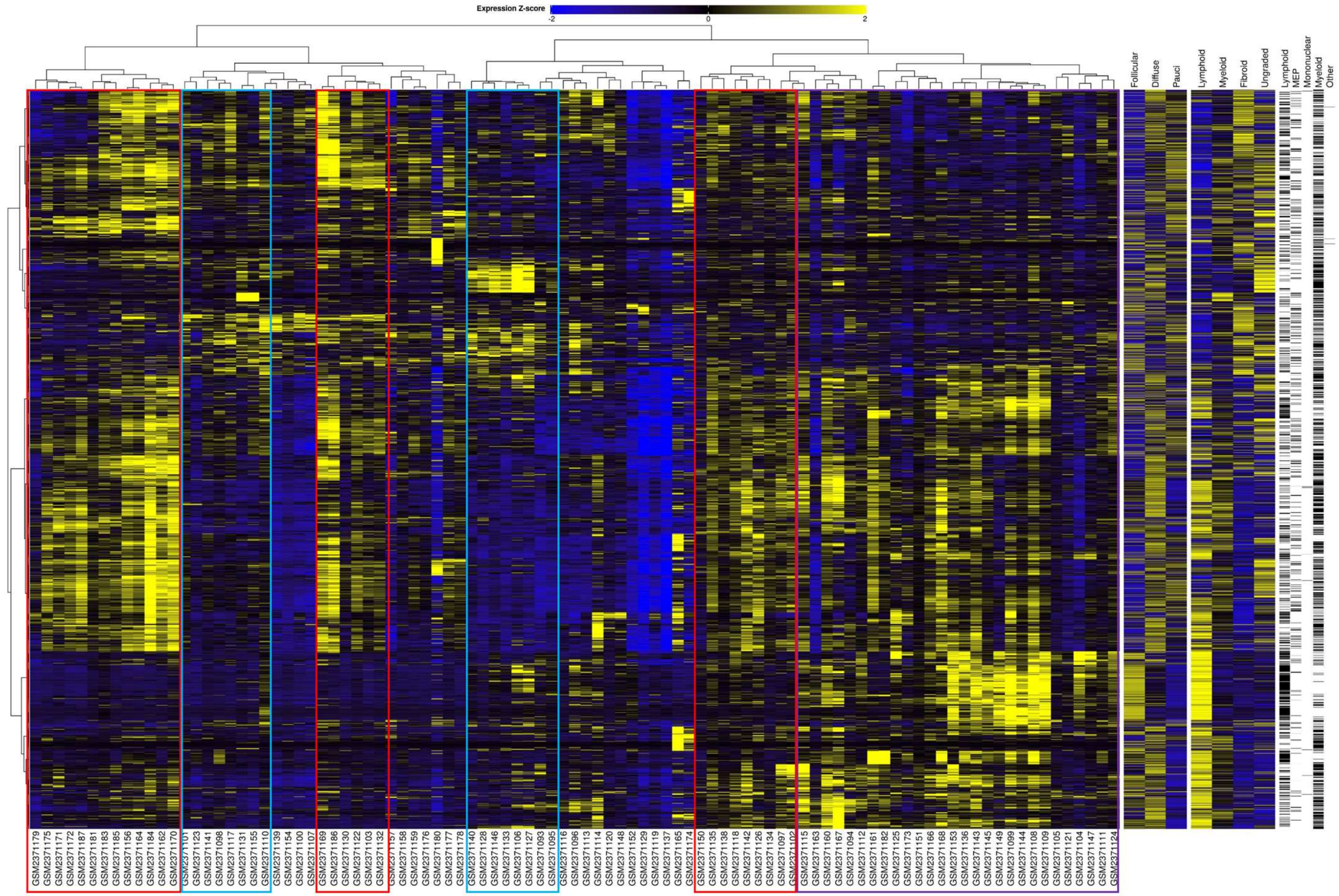


Figure 6.4: Immune and stromal markers in established rheumatoid arthritis.

Patients with established disease from GSE89408 are contrasted with the archetypes derived from GSE48780 (Follicular, Diffuse, Pauci) and E-MTAB-6141 (Lymphoid, Myeloid, Fibroid, Ungraded). This dataset lacks metadata to define pathology therefore samples were characterised by similarity to the archetypes. Samples that are representative of these pathologies are denoted by the coloured box (Follicular – purple; Diffuse – red; Pauci – cyan) which was then utilised in further analysis.

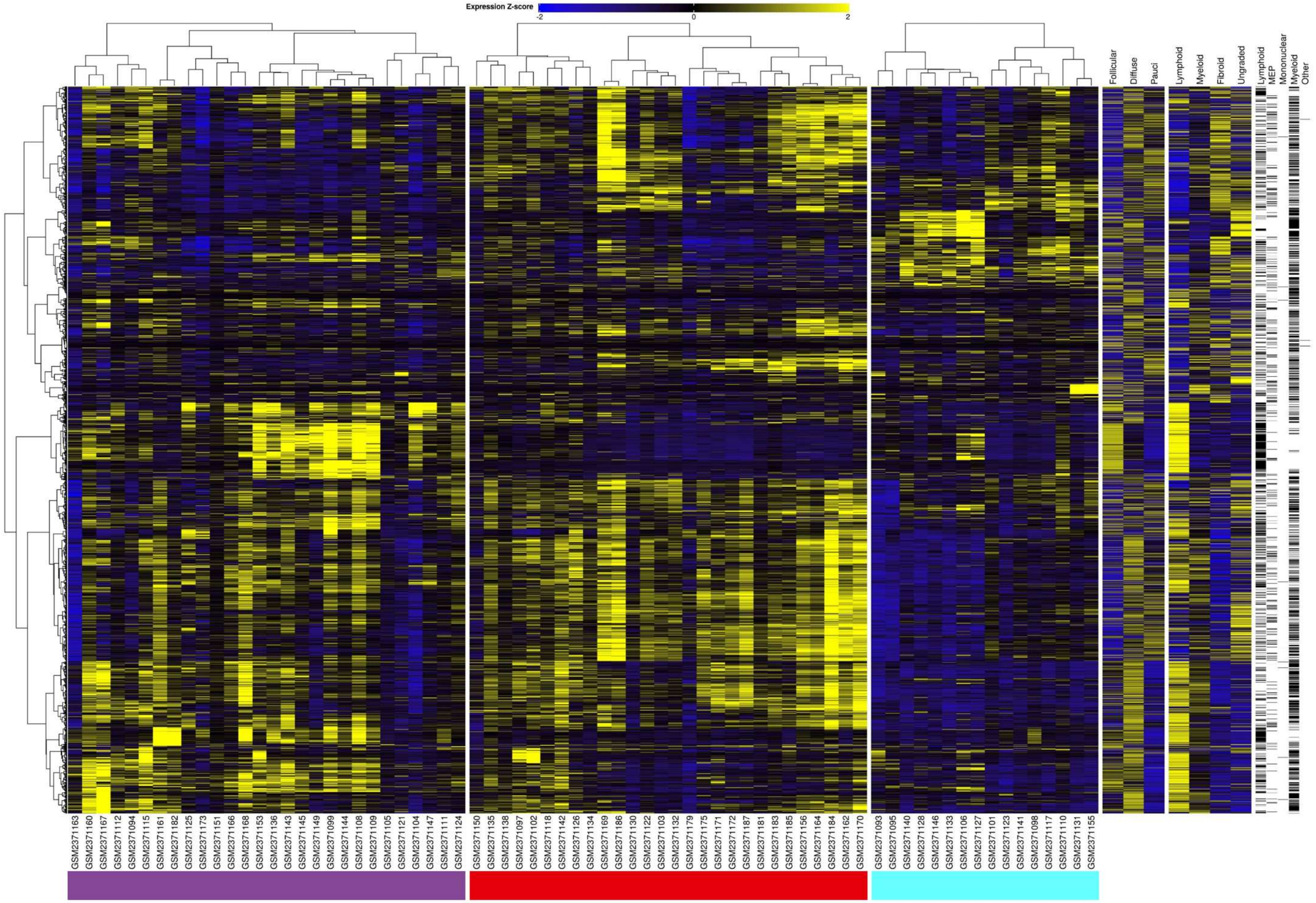


Figure 6.5: Immune and stromal markers in established arthritis with defined pathologies.

The defined samples that show clear stratification of established rheumatoid arthritis patients (GSE89408) that match the immune profile of the archetypes. Resulting in 22 follicular, 7 diffuse, and 13 pauci-immune samples that can be tested going forwards.

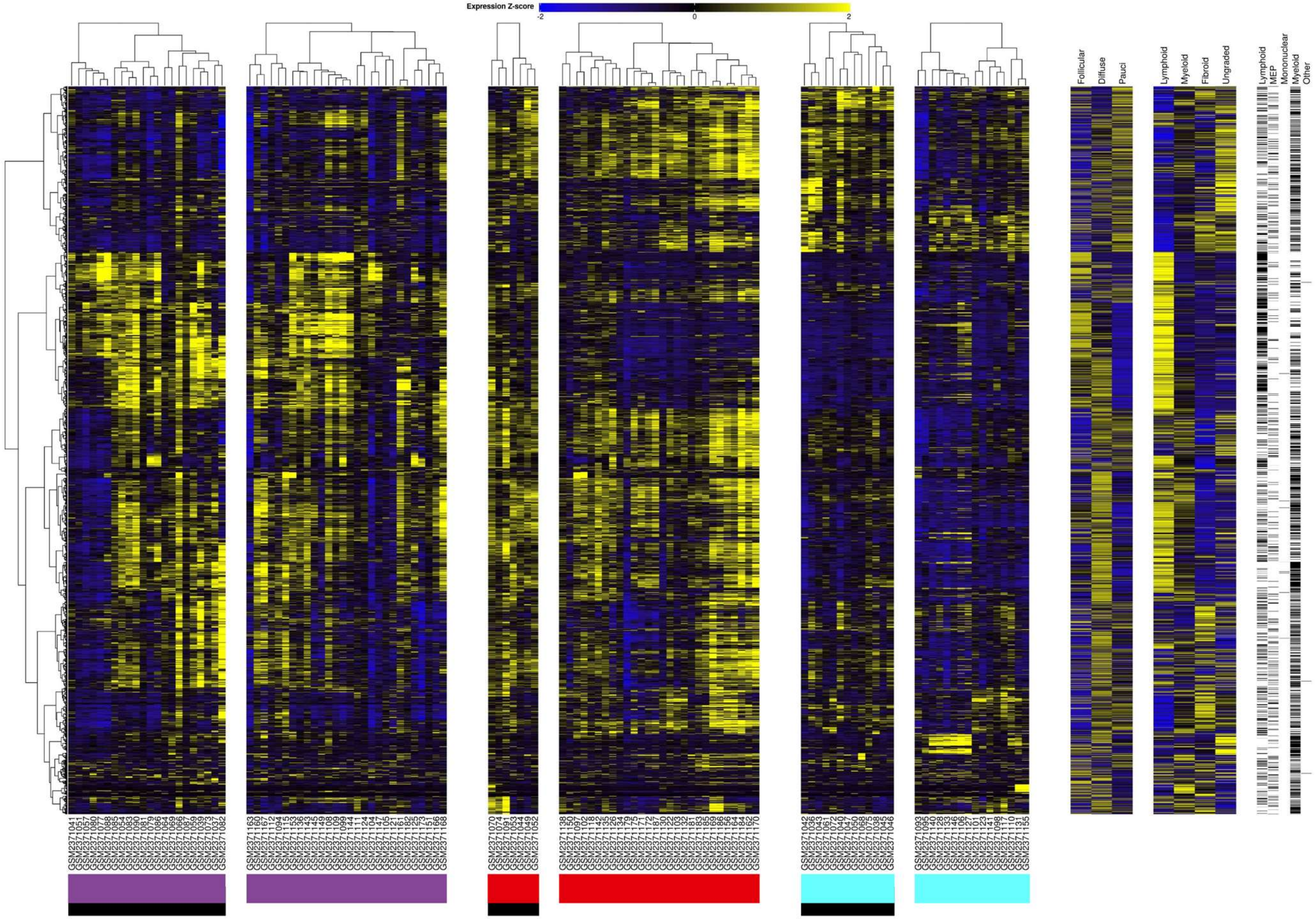


Figure 6.6: Comparing the immune and stromal transcriptomes of early and established disease.

Pathotypes are grouped together to contrast early-stage with that of established disease. Follicular and Pauci pathologies in both early and established show a strong correlation with that of the archetypes derived from both GSE48780 and the PEAC datasets. Diffuse samples have a mixed expression during the early stage and resemble the archetypes in established, although in this dataset with an overlap with markers associated with Pauci-immune.

6.5.2. Stratified early and established disease.

Figure 6.2 depicts the entire series of early-stage samples that show characteristic immune and stromal cell profiles associated with the archetypical expression of each pathotype. To enhance the interpretation of these data, individual transcriptomic datasets were grouped according to a defined form of pathology (Figure 6.3). Data acquired from biopsies with early stage synovitis resulted in 21 follicular, 7 diffuse, and 13 pauci-immune samples. This approach was repeated for data acquired from samples with established disease (Figure 6.4 and condensed down in Figure 6.5). These resulted in the classification of 28 follicular, 28 diffuse, and 16 pauci-immune forms of synovitis. These groupings were taken forward for further performance testing of the prediction tools.

As discussed in the previous chapter, both the follicular and pauci-immune pathologies show clear similarities in the immune and stromal markers between early-stage and established disease. However, as previously discussed, when looking at the diffuse pathology (Figure 6.6), early-stage samples demonstrated a mixed expression of cellular markers that don't clearly identify with a specific archetypal signature for an individual with this form of diffuse synovitis.

6.5.3. Testing classifier performance in early and established disease.

Next, to understand how well these signatures discriminate the pathologies in both early and established disease, the classifiers were assessed in the stratified samples explored in 6.5.2.

Employing the classifiers generated from the analysis of established disease (derived in Chapter 5) with the ones generated from early-stage synovitis (derived here from the PEAC dataset), experiments tested the utility of these tools as predictors of disease.

Classifier performance was tested using the pathologies assigned by immune and stromal transcriptional profiles in the previous section. Performance of the classifiers was assessed using the area under the curve and graded 'good', 'acceptable', and 'poor' based on guidelines outlined in the literature(179,180) (≥0.8, ≥0.6<0.8, <0.6 respectively).

Investigating the early-stage samples in Figure 6.7 shows the performance of the three best models identified previously (**A**: 13 gene, **B:** 17 gene, and **C** Random Forest), and the best model derived from the PEAC dataset (**D**: Fit123). As shown in the previous chapter, these three models continue to demonstrate good performance at discriminating the follicular and pauci-immune pathotypes, and also shows acceptable performance with the diffuse pathology. Unfortunately, the best model from the PEAC dataset does not replicate this performance; whilst demonstrating good performance for follicular, it fails to stratify the diffuse and pauci-immune pathotypes. The performance of all applicable models discussed in this and previous chapters are outlined in Table 6.3.

For the samples with established disease (Figure 6.8), reduced performance is observed in all of the -fit models. All of the models demonstrated good or acceptable performance in discriminating the follicular pathology. However, whilst these models identified in chapter 5 performed well at stratifying pauci-immune samples in early disease, this was reduced for pauci-immune samples in established disease. The signatures generated in this chapter have no ability to stratify pauci-immune samples in established disease.

Table 6.3: Classifier performance as assessed by AUC when stratifying early-stage and established rheumatoid arthritis patients from GSE89408.

Colours represent good (green≥0.8), acceptable (yellow≥0.6<0.8) and poor (red, <0.6) performance at stratifying the pathologies

Model			Early-Stage			Established		
			Follicular	Diffuse	Pauci	Follicular	Diffuse	Pauci
BootsPLS	GSE48780	13 gene	0.87	0.61	0.99	0.76	0.48	0.51
		17 gene	0.82	0.67	0.99	0.82	0.57	0.61
		fit1556	0.93	0.65	1.00	0.82	0.51	0.62
		fit2078	0.93	0.65	1.00	0.82	0.51	0.62
		fit4657	0.93	0.65	1.00	0.82	0.51	0.62
		fit5184	0.93	0.65	1.00	0.82	0.51	0.62
		fit5237	0.95	0.68	1.00	0.82	0.53	0.61
		fit5495	0.93	0.65	1.00	0.82	0.51	0.62
		fit6104	0.93	0.65	1.00	0.82	0.51	0.62
		fit6747	0.93	0.65	1.00	0.82	0.51	0.62
		fit8395	0.95	0.67	1.00	0.84	0.57	0.57
		fit8913	0.93	0.65	1.00	0.82	0.51	0.62
		fit8945	0.93	0.65	1.00	0.82	0.51	0.62
		fit4948	0.92	0.64	0.99	0.81	0.51	0.58
	PEAC	fit22	0.68	0.32	0.33	0.81	0.23	0.37
		fit42	0.72	0.32	0.33	0.81	0.22	0.38
		fit123	0.92	0.31	0.50	0.85	0.19	0.41
		fit1234	0.72	0.33	0.34	0.81	0.23	0.41
		fit2020	0.72	0.32	0.33	0.81	0.22	0.38
RandomForest			0.80	0.73	0.97	0.71	0.59	0.79

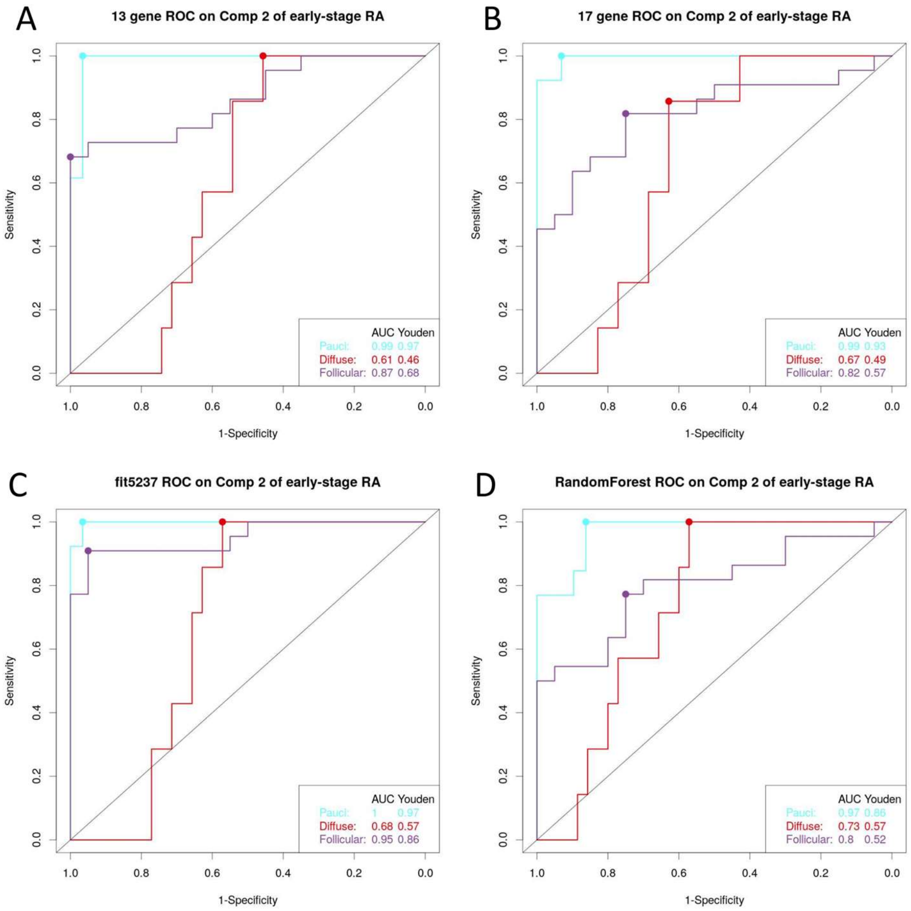


Figure 6.7: Testing classifier performance against early-stage disease.

Stratified early-stage rheumatoid arthritis samples as defined in Figure 6.3 were tested using previously determined models (A-C) and the new gene signature derived from the PEAC dataset (D). Models 13 gene, 17 gene, and RandomForest show good performance stratifying Pauci (Cyan) and Follicular (Purple) pathologies, in addition to reasonable performance in Diffuse (Red). The new signature (D) shows good performance with the Follicular pathology but fails to discriminate Pauci and Diffuse.

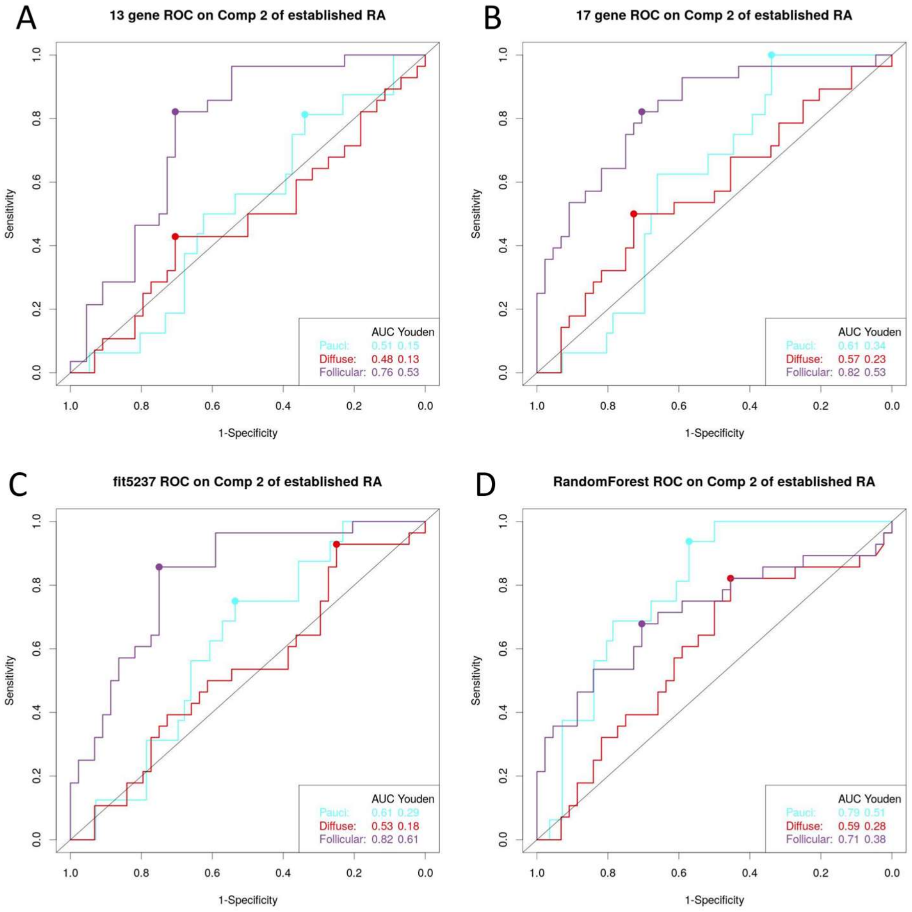


Figure 6.8: Testing classifier performance against established disease.

Stratified established rheumatoid arthritis samples as defined in Figure 6.6 were tested using previously determined models (A-C) and the new gene signature derived from the PEAC dataset (D). Models 13 gene, 17 gene, and RandomForest show good performance stratifying the Follicular (Purple) pathology, but limited performance with the Pauci (Cyan) and Diffuse (Red). The new signature performs well for Follicular but fails to discriminate Diffuse and Pauci pathologies.

6.6. Discussion

Experiments outlined in Chapter 5 demonstrated that computational tools designed to classify synovial pathology based on transcriptomic data could be used to distinguish pauci-immune and follicular synovitis. However, this approach was unable to accurately predict the identification of myeloid-rich or diffuse synovitis. The analysis presented in this Chapter reinforces this view and shows that Classifier tools are poor predictors of myeloid-rich pathology in datasets derived from both early stage synovitis and late-stage or established disease. Based on this analysis, I conclude that myeloid-rich synovitis may be sub-classed into pathologies that display transcriptional profiles that veer more towards either pauci-immune or follicular synovitis. Thus, making the classification more challenging.

The limited forms of metadata associated with the GSE87408 dataset meant that the nature of the synovial pathology had to be inferred through comparisons with transcriptomic datasets obtained from immune and stromal cells. The validity of this type of approach was confirmed when viewing datasets displaying follicular-like or pauci-immune-like gene signatures and the identity of these forms of pathology were clearly delineated in early and established forms of disease. However, the classification of diffuse synovitis was again difficult to predict.

According to the literature, there is little difference between early and established disease. Histological examinations revealed no differences with regards scoring of cellular infiltrates and hyperplasia(189,190) or the presence of molecules such as cytokines and matrix metalloproteins(188). This is replicated when exploring the transcriptomes, with Figure 6.6 showing the stability of the follicular and pauci-immune transcriptional profiles but highlights the difference between early-stage and established for those diffuse patients. The overlap observed in early diffuse samples may well reflect that early in disease there is a distinct and transient change to the cytokine profile of the synovial fluid(191). This change has been associated with cytokine production by T-cells and stromal-cells, which may explain the “muddy’ profile observed.

The signatures derived from the PEAC dataset resulted in a completely new set of genes that perform well at predicting follicular pathology in both early and established disease. Unfortunately, these new signatures don’t exhibit improved performance with regards to the diffuse pathology and the classifier was less specific for detection of pauci-immune synovitis.

In early disease, all of the models derived in the previous chapter exhibited good or excellent performance for the follicular and pauci-immune and demonstrated some acceptable ability for the diffuse samples. The best overall model in early disease would be fit5327, a classifier based on 36 genes, which exhibited near-perfect stratification of the follicular and pauci-immune samples, but also good performance with the diffuse samples. The RandomForest model showed the best performance for classifying diffuse, but at the cost of slightly reduced performance with the follicular pathology.

In established disease, however, all models showed reduced ability to stratify disease. RandomForest continues to exhibit the best performance with the diffuse pathology, but not enough to be considered acceptable. It does, however, demonstrate the best performance for classifying the pauci-immune samples.

Overall, these models demonstrate that it is possible to predict follicular and pauci-immune at an early stage of the disease. Therefore, combined with the knowledge of pathology effect on therapeutic response (for example reduced response to TNFα blockade in pauci-immune(72)) this opens the door towards a more effective treat-to-target approach in early treatment, which is essential for improving clinical outcomes.

7. Monitoring clinical responses to therapeutic interventions.

7.1. Introduction.

Treatment guidelines for rheumatoid arthritis follow a prescribed regimen of disease-modifying anti-rheumatic drugs (DMARD's) or biologic agents (Figure 5.1). While the adherence to these guidelines has dramatically improved patient outcomes, many patients fail to respond to conventional DMARD's or biological drugs. Results from the previous chapters have illustrated the heterogeneous nature of synovitis in RA. It is, therefore, proposed that alternate patterns of inflammatory regulation shape the course of disease progression. These differences are likely to influence the way a patient or patient group respond to therapy.

Following diagnosis, the ACR and EULAR guidelines recommend that patients immediately start on methotrexate (MTX) as the initial disease-modifying therapy(49). For patients with prognostic factors indicative of high disease activity, MTX may be combined with either glucocorticoids or some other DMARD (e.g. sulphasalazine, leflunomide, hydroxychloroquine). However, only 30% of patients achieve low disease activity following MTX monotherapy (51), necessitating the use of a biological drug.

First line biological drugs include anti-TNF therapies such as adalimumab, etanercept, and infliximab. These therapies are similar in terms of efficacy and tolerability and target the TNFα cytokine(192). Although these therapies have revolutionised the treatment of immune-mediated inflammatory diseases, approximately 25-35% of patients with RA fail to respond to anti-TNF therapy. For these patients, biological drugs that target other inflammatory pathways are frequently prescribed. For example, monoclonal antibodies that inhibit IL-6R signalling (e.g., tocilizumab; TCZ), deplete circulated CD20^{+} B-cells (e.g. rituximab; RTX), or small molecule inhibitors that block specific intracellular signalling pathways (e.g., tofacitinib).

One of the surprising things about many of these therapies is that their mechanism of action in RA is still relatively unclear (74,193). For example, whilst MTX is known to inhibit dihydrofolate reductase and prevents thymidine synthesis in cancer(194), its role in RA is less clear. As in cancer, T-cell proliferation is controlled by MTX; however, it is also involved in interrupting cell signalling, changes in cell adhesion molecules, and cellular migration(193). Similarly, whilst depletion of autoantibody producing B-cells is a significant mode of action of RTX, its clinical efficacy in B-cell poor pauci-immune patients indicates that the therapeutic mode-of-action of RTX in RA maybe broader than B-cell depletion (62,90,145,195).

Numerous clinical studies have interrogated the effect of these therapeutics on the synovium (71,81,82,86,97,144,145,196) to understand pathological mechanisms within the inflamed joint. However, many of these studies have ultimately been confounded by a lack of detailed information specific the precise clinical presentation of synovitis seen in patients. Adopting the classifier tools developed in the previous Chapters, experiments outlined here tested whether these tools could be used to identify biological responses to therapeutic intervention specific to the individual pathologies. Results obtained through this study showed that synovial histopathology had to be taken into account when predicting response to therapy in RA(81). Moreover, my findings show that commonly prescribed biological drugs modulate different sets of genes in different forms of synovitis.

7.2. Hypothesis.

Experiments will evaluate the overarching hypothesis that patient stratification based on synovial histopathology will increase the ability to detect the mechanisms of action of biological therapeutics. Thus, allowing the identification of signatures that predict response to an individual biological drug therapy.

7.3. Aims.

To address this hypothesis, the bioinformatic prediction tools developed in the previous Chapters were used to: (1) classify synovial pathology in absence of reliable histological data, and (2) identify differential patterns of gene regulation that predict a therapeutic response to treatment. Specifically, open access repository datasets were used to evaluate treatment responses to MTX, RTX and TCZ .

7.4. Materials and Methods.

7.4.1. Repository datasets and associated metadata.

Several previously described datasets were used. The specifics for each of these datasets are summarised in Table 7.1. For this chapter, GSE24742 (containing synovial transcriptomic data from patients before and after RTX treatment) and GSE45867 (containing synovial transcriptomic data from patients before and after MTX or TCZ treatment) datasets were interrogated.

GSE24742 provides a wealth of metadata, including histological scores relating to the synovial pathology and specific immune cell scores based on immuno-histochemical staining. Using the rules outlined in Chapter 5.4.1, synovial joint pathology was defined using this metadata. GSE45867 provides less information but contains DAS28 scores as a response to therapy.

1.1.1. Pathotype stratification.

Samples before administration of MTX or TCZ (GSE45867) were compared with the archetypes derived in Chapter 4.

For both MTX and TCZ treatments, samples were clustered according to the 13 gene signature (1-Spearman's correlation and Ward's D) generated in Chapter 4.5.6. This form of analysis was initially applied to datasets obtained before treatment and subsequently extrapolated to the datasets extracted from sample post treatment. Thus, allowing a direct comparison of data before and after intervention.

1.1.2. Differential expression.

Differential gene expression was determined using the limma package. To simplify the interpretations, samples were defined as responders ("good" and "moderate") or non-responders ("poor") using the DAS28 response criteria(43). Thus, allowing changes in gene expression as a response to therapy to be compared with improvements in DAS28 scores. Here, the sub-classification of transcriptomic data according to synovial pathology was used to understand how patients with defined forms of pathology respond to MTX, RTX or TCZ intervention.

1.2. Results.

1.2.1. Response to therapy and pathotype stratification

Inspection of the GSE24742 dataset identified 24 samples from 12 patients taken before and 12 weeks after RTX treatment. Further analysis of these 12 patients revealed 5 patients with follicular pathology, a further 6 patients with a diffuse form of synovitis, and 1 with pauci-immune synovitis. Of these patients, 9 responded to therapy and 3 displayed no improvement in disease activity (Figure 7.1).

The GSE45867 sample cohort contained 40 samples from 20 patients (12 received TCZ, 8 received MTX). The analysis presented in Figure 7.2 shows that this dataset could be delineated into patients with follicular (comprising 4 patients), diffuse (comprising 9 patients) or pauci-immune (comprising 7 patients) synovitis.

Whilst the available data from GSE24742 and GSE45867 allowed the gathering of information on a patient's responses to treatment, the limited numbers of patients characterised as non-responders restricted the ability to understand the mechanisms that may prohibit effective therapy (Table 7.2).

Table 7.1: Summary of the number of patients belonging to each group defined by pathology and response to therapeutics.

Dataset	Treatment	Responders			Non-Responders		
		Follicular	Diffuse	Pauci	Follicular	Diffuse	Pauci
GSE24742	RTX	4	3	1	1	3	NA
GSE45867	MTX	1	2	1	1	1	2
	TCZ	1	6	4	1	NA	NA

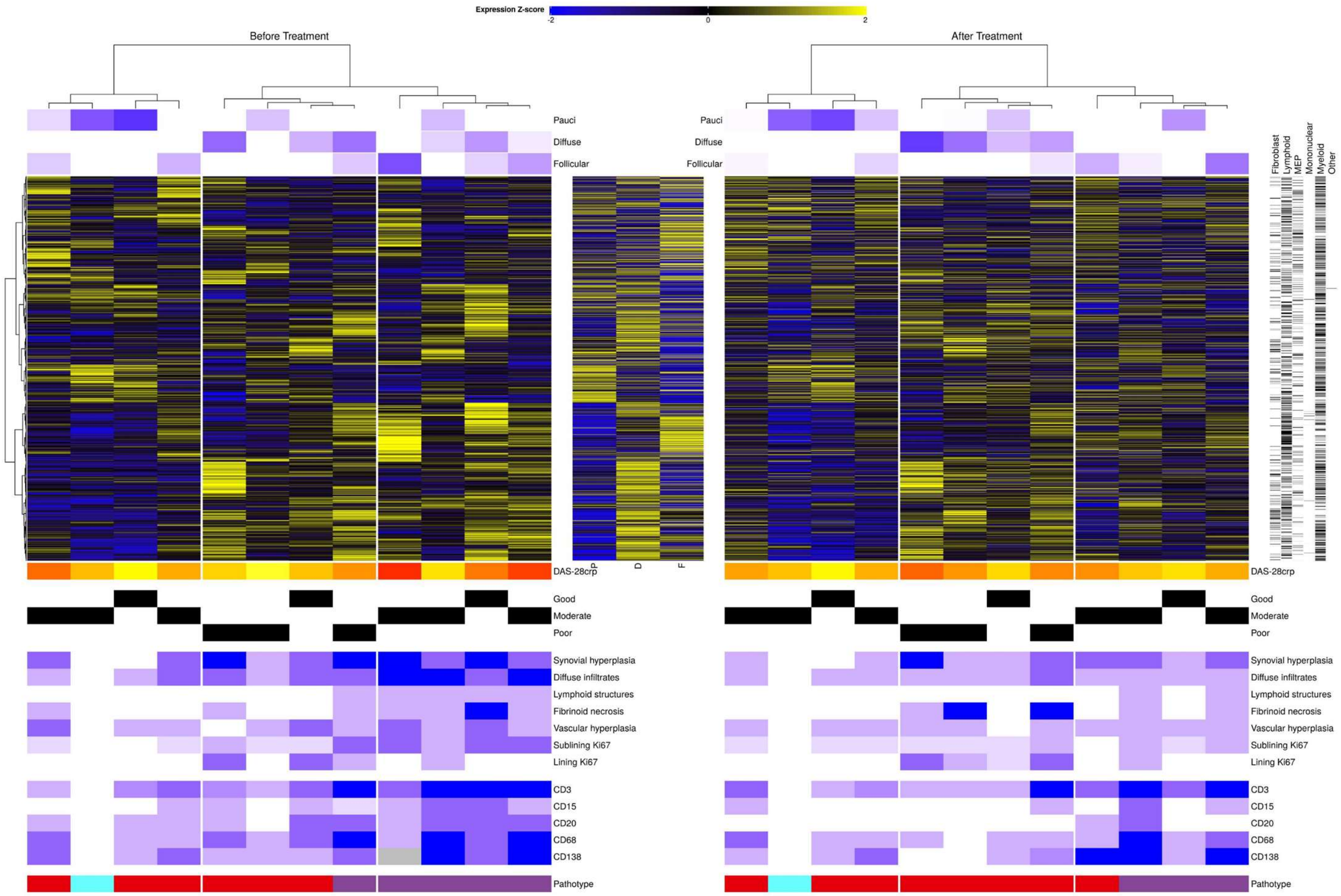


Figure 7.1: Changes in immune and stromal markers as a result of rituximab therapy.

Immune and stromal markers for GSE24742, looking at response to Rituximab before and after treatment. Before treatment samples were clustered using the 13 gene signature (1-Spearman correlation and Ward's D), and compared to the archetype expression (Pearson's correlation) shown as the top annotation of the heatmap. GSE24742 contains comprehensive histology metadata, the scoring for the relevant features are shown under the heatmaps. Pathotype is defined using the rules outlined in Table 5.1, leading to 5 Follicular, 6 Diffuse, and 1 Pauci-immune sample.

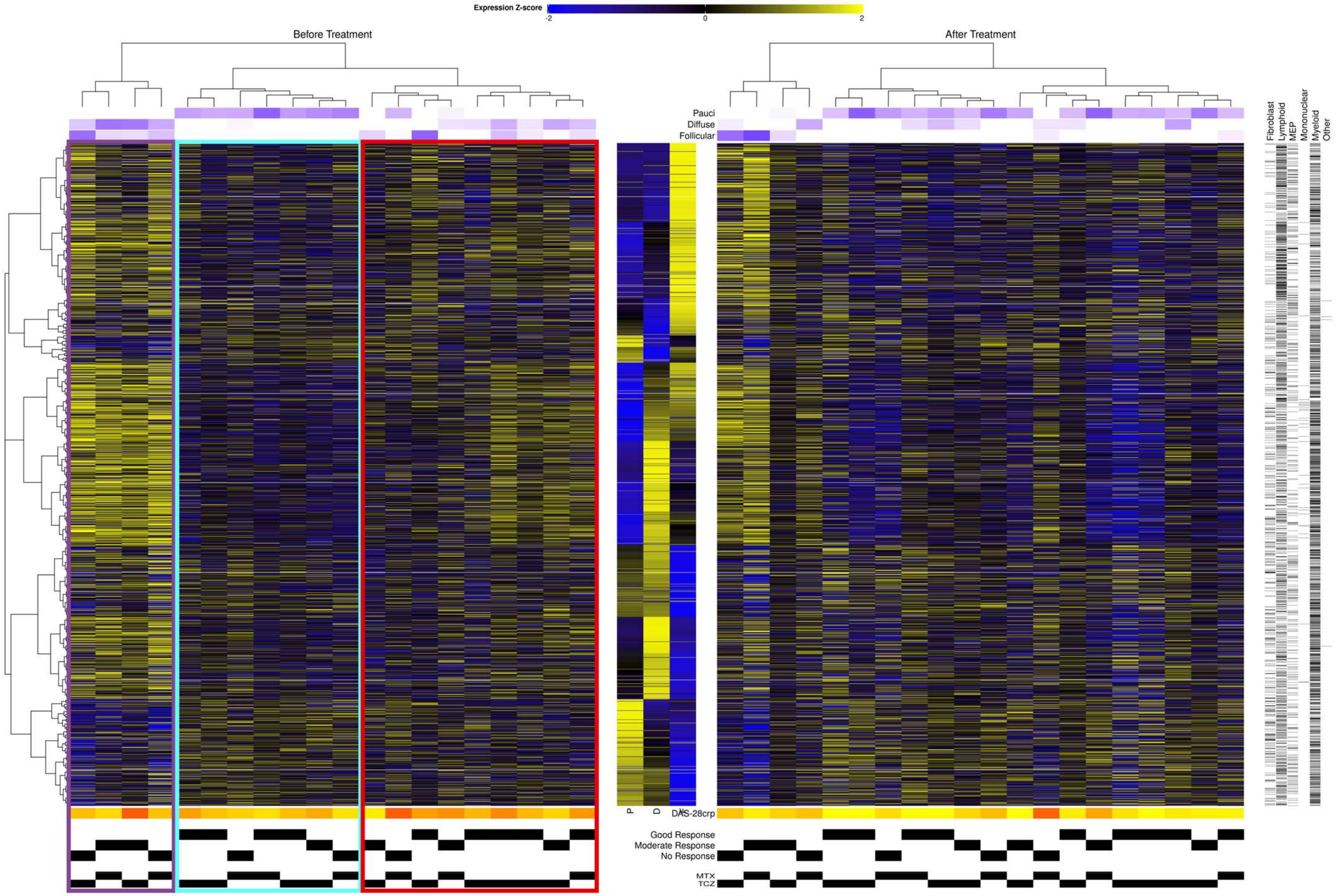


Figure 7.2: Immune and stromal markers for GSE45867, looking at response to Methotrexate or Tocilizumab before and after treatment.

Before treatment samples were clustered using the 13 gene signature (1-Spearman correlation and Ward's D) and compared to the archetype expression (Pearson's correlation) shown as the top annotation of the heatmap. Pathotypes were assigned as by this correlation with archetype, as indicated with the boxes around the samples; 4 Follicular samples (purple), 9 Diffuse (red), and 7 Pauci-immune (cyan).

Table 7.2: Publicly available datasets that specifically explore the response to therapeutics in RA synovium.

Accession number	Platform	# samples	Description
GSE45867	Affymetrix HGU133plus2	40	Synovial biopsies before and after treatment with either Tocilizumab (12 patients: 24 samples) or Methotrexate (8 patients: 16 samples)
GSE24742	Affymetrix HGU133plus2	24	Synovial biopsies before and after treatment with Rituximab (12 patients)
GSE15602	Affymetrix HGU133plus2	11	Synovial biopsies after treatment with Adalimumab
E-TABM-104	KTH H. sapiens 29.8k cDNA v2/KTH H. sapiens 30.5k cDNA array v1	32	Synovial biopsies before and after treatment with infliximab (10 patients)
GSE21537	KTH H. sapiens 30.5k cDNA array v1	62	RA synovial biopsies before infliximab treatment
GSE97165	Illumina HiSeq 2000	38	Synovial biopsies before and after treatment with triple DMARDs (19 patients)

1.2.2. Differentially expressed genes.

Whilst single samples are not desirable, the limma package allows the incorporation of information from other samples to be explored(9). Here, it is assumed that variance is the same between groups and maybe simulated from a single sample. Despite this limitation in the analysis, stratification still provided valuable information on the unique properties of the pathology and the response to therapy.

Table 7.3: Number of significantly differentially expressed probe-sets associated with therapeutic response in the synovium of RA patients for the two datasets explored.

Dataset	Treatment		Responders			Non-Responders		
			Follicular	Diffuse	Pauci	Follicular	Diffuse	Pauci
GSE24742	RTX	Unadjusted (P ≤0.01)	670	101	666	142	132	NA
		Adjusted (FDR) (P ≤0.05)	0	0	24	0	0	NA
GSE45867	MTX	Unadjusted (P ≤0.05)	8592	5954	1606	9552	1813	1628
		Adjusted (FDR) (P ≤0.05)	593	13	2	1301	0	0
	TCZ	Unadjusted (P ≤0.05)	1641	6538	1616	2263	NA	NA
		Adjusted (FDR) (P ≤0.05)	5	548	0	20	NA	NA

1.2.2.1. GSE24742 – Increased detection of differentially expressed genes as a response to rituximab.

Analysis of the GSE24742 dataset by the authors of the original publication identified 549 probe-sets that were differentially expressed as a response to RTX treatment (145). However, their statistical analysis (P ≤0.01) was conducted on unadjusted p-values. This type of analysis seemed inappropriate and the identification of 549 probe-sets from the 54,765 probe-sets within the microarray platform would be expected to arise through pure chance based on this statistical method. I therefore performed a revised analysis of these datasets.

Here, reanalysis found that the classification of pathology substantially improved the ability to distinguish a specific set of differentially expressed genes that remain significant when adjusting of multiple testing. Ultimately, the small size of the dataset means that there is limited power to detect effects, and for most conditions fails to detect any significantly differentially expressed probe-sets after multiple testing correction. Table 7.3 outlines how incorporating the stratification of pathology and response results in an increase in the number of probe-sets that are significant at the same level as the original publication – although this is based on raw unadjusted p-values.

Due to the limited power to detect significantly differentially expressed genes due to the small sample size, further analyses were done using the unadjusted p-values. Figure 7.3 demonstrates the differentially expressed genes identified for all the conditions (pathotype and response), annotating the most significant probe-sets.

By contrast, incorporation of response and pathology increases the number of probe-sets observed as significantly differentially expressed, with some still significant after multiple testing (24 probe-sets in pauci-immune responders). This more detailed interpretation was missed in the original paper (the most significant result being 1.31E-4, unadjusted). Looking across all the pathologies and responses results in 1'659 probe-sets that are significant at 0.01 unadjusted, considerably more than the original publication. Figure 7.3 shows volcano plots of all the conditions measured, highlighting those genes with the most significant results.

Molecular pathway analysis of the differentially expressed genes identified several biological pathways involved in each of the synovial pathologies and showcased several processes that were selectively blocked by therapy. A summary of these findings is shown in Table 7.4.

Table 7.4: Number of genes that are significantly differentially expressed (P ≤0.05, non-adjusted), and pathways associated with these genes, following rituximab treatment.

Condition	Upregulated genes (probes)	Downregulated genes (probes)	Associated pathways
Follicular Responders	1038 (1706)	1394 (1700)	TLR signalling, IL-2 signalling, Jak-STAT signalling, cytokine signalling
Follicular Non-Responders	418 (555)	367 (472)	PI3 kinase and angiotensin II signalling
Diffuse Responders	285 (433)	402 (498)	Wnt signalling
Diffuse Non-Responders	292 (345)	546 (781)	FGFR3 and IL-6 signalling
Pauci-immune Responders	1029 (1439)	1016 (1293)	Fatty acid metabolism, EGF receptor signalling

1.2.2.2. GSE45867

1.2.2.2.1. Methotrexate.

The original investigation(144) of the effects of methotrexate on the synovium identified 1'196 probes as significantly differentially expressed (P ≤0.05) without multiple testing correction. Incorporation of the pathotypes and response to therapy not only increased the number of probe-sets, but also identified more probe-sets that are significant after multiple testing correction. Table 7.5 breaks down the 1'738 probes that are significantly differentially expressed after multiple testing correction (P ≤0.05, after FDR correction). Figure 7.4 highlights some of the most significant genes for each condition.

Table 7.5: Number of genes that are significantly differentially expressed (P ≤0.05, FDR adjusted), and pathways associated with these genes, following methotrexate treatment.

Condition	Upregulated genes (probes)	Downregulated genes (probes)	Associated pathways
Follicular Responders	257 (402)	307 (338)	BMP signalling, IGF regulation, mTOR signalling
Follicular Non-Responders	572 (724)	477 (577)	Spliceosome, mRNA processing, autophagy
Diffuse Responders	2(3)	6 (9)	NF-κB signalling, TLR signalling
Pauci-immune Responders	-	2 (2)	-

1.2.2.2.2. Tocilizumab.

The original paper describing the effects of Tocilizumab on the synovium provided a large number (6,683) of significant probe-sets (144) Again, this number of probe-sets was derived without multiple testing correction. Without the raw P-values (supplementary table 2 from the original publication lists the probe-sets, but not p-values) it is not clear how many of these would be significant after correction. Incorporation of the pathotype and response increased the number of significantly differentially expressed probes detected to 10,660, but this is before multiple testing correction. When adjusted using FDR, 569 probe-sets are identified as significantly differentially expressed and are explored in Table 7.6. Figure 7.5 highlights the most significant differentially expressed genes.

Table 7.6: Number of genes that are significantly differentially expressed (P ≤0.05, non-adjusted), and pathways associated with these genes, following tocilizumab treatment.

Condition	Upregulated genes (probes)	Downregulated genes (probes)	Associated pathways
Follicular Responders	1 (1)	4 (4)	IL-1 regulation of extracellular matrix, osteopontin/osteoclast signalling
Follicular Non-Responders	-	18 (20)	Tryptophan catabolism, nitric oxide response
Diffuse Responders	43 (50)	364 (498)	Netrin-1 signalling, Wnt signalling, PhopholipaseD signalling, IL-2 signalling, chemokine signalling
Pauci-immune Responders	-	-	-

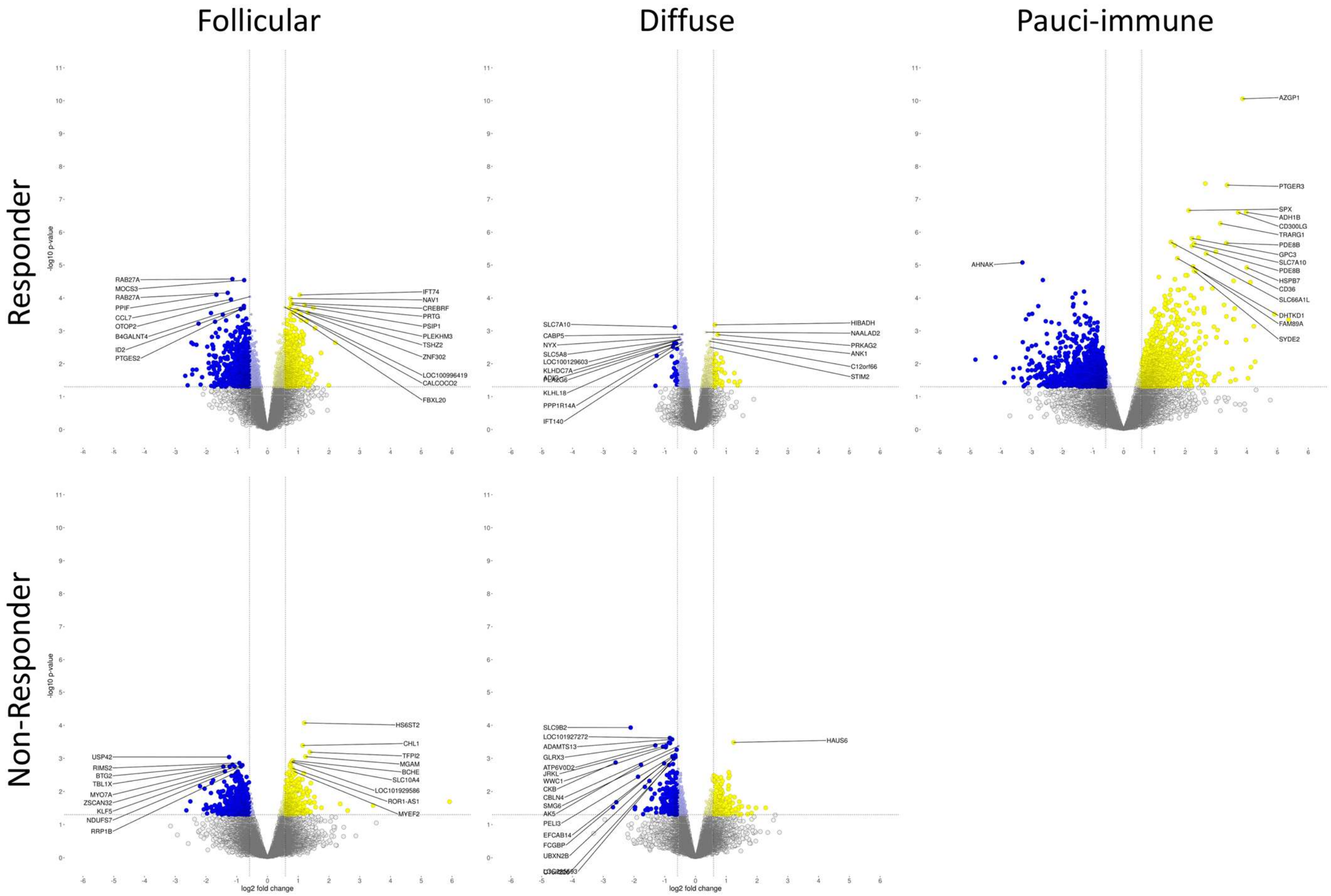


Figure 7.3: Differentially expressed genes for the three pathotypes as a result of Rituximab treatment.

Samples from GSE24742 were stratified to pathology and tested for differential expression between baseline and 12 weeks after treatment for responders and non-responders. Due to the small sample size only a small number of genes are significant after multiple testing correction (FDR) in pauci-immune responders (24 genes) but not other pathologies/responses, therefore unadjusted p-values are plotted. Yellow and blue points represent significant genes (P ≤0.05) and more than 1.5-fold change – up and down respectively.

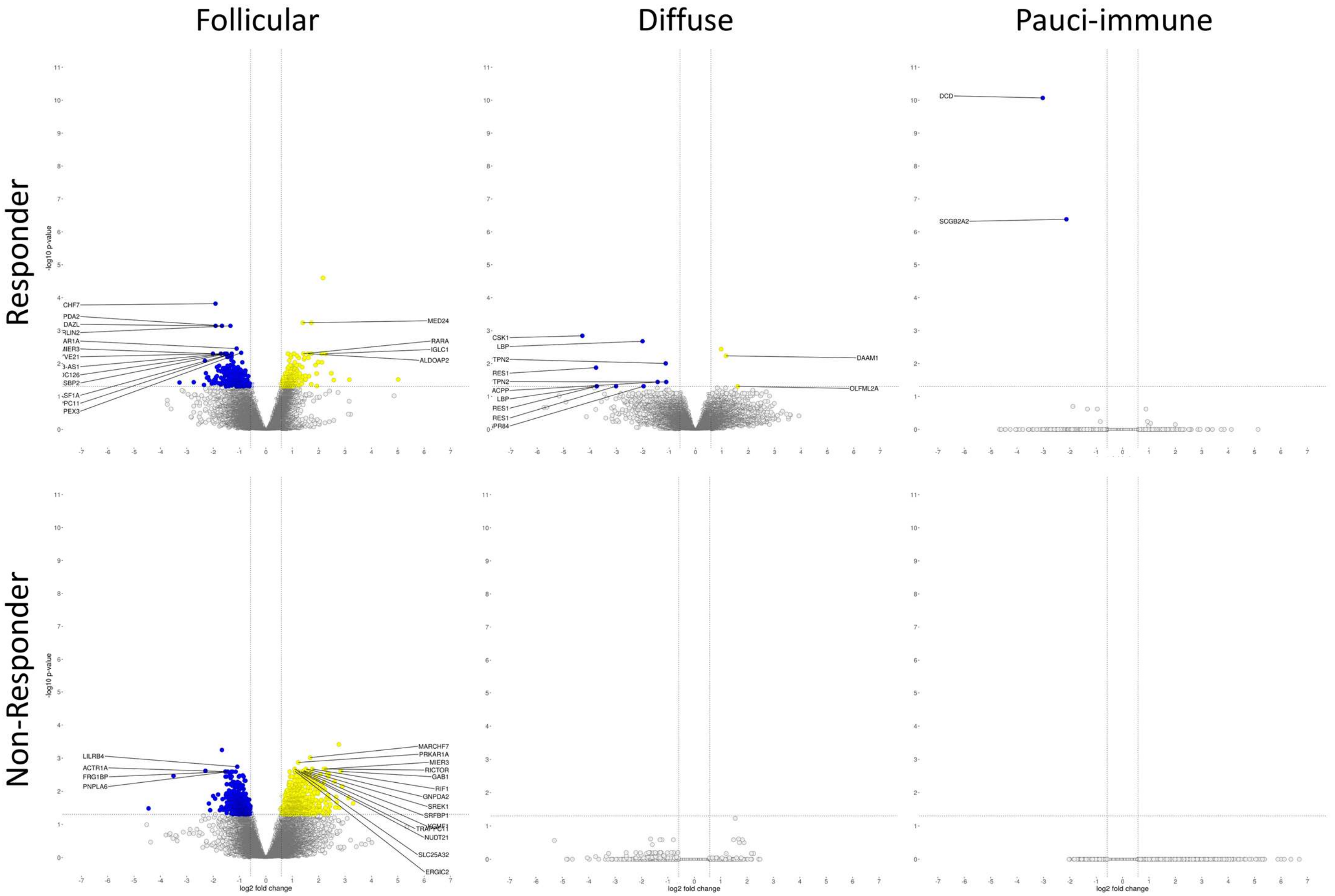


Figure 7.4: Differentially expressed genes for the three pathotypes as a result of Methotrexate treatment.

Samples from GSE45867 were stratified to pathology and tested for differential expression between baseline and 12 weeks after treatment for responders and non-responders. X axis shows fold change, and y axis –log10(adjusted P-value (FDR)) Yellow and blue points represent significant genes (P ≤0.05) and more than 1.5-fold change – up and down respectively. 20 most significant genes (below 0.05 threshold) are annotated

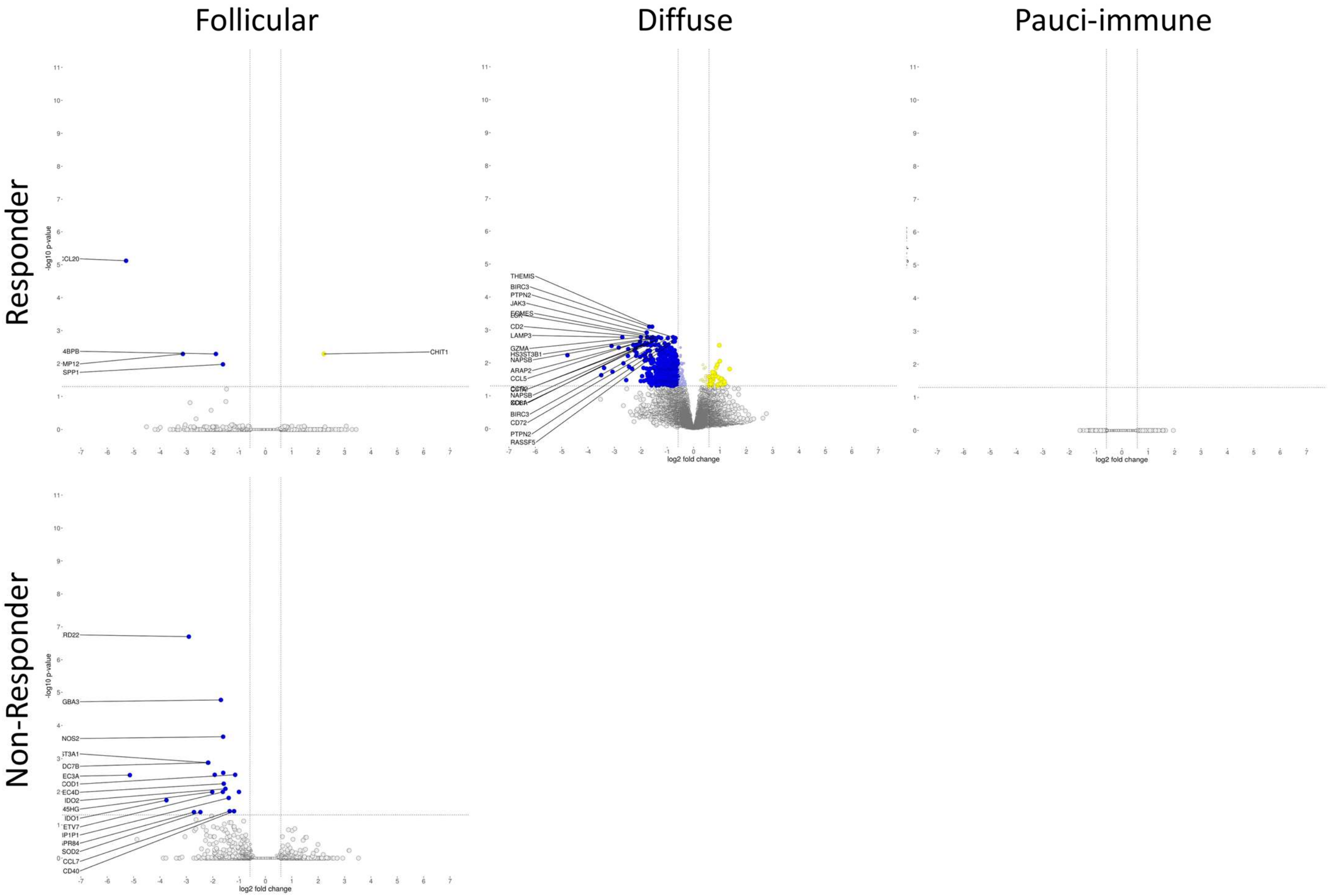


Figure 7.5: Differentially expressed genes for the three pathotypes as a result of Tocilizumab treatment.

Samples from GSE45867 were stratified to pathology and tested for differential expression between baseline and 12 weeks after treatment for responders and non-responders. X axis shows fold change, and y axis –log10(adjusted P-value (FDR)) Yellow and blue points represent significant genes (P ≤0.05) and more than 1.5-fold change – up and down respectively. 20 most significant genes (below 0.05 threshold) are annotated

1.3. Discussion

Results presented in this chapter has demonstrated that stratification of patients improves the ability to detect differentially expressed genes as a response to biological drug therapy and provides new opportunities to understand the pathways that drive pathology. What these datasets have highlighted predominantly is the issue with small sample sizes, especially when not pre-screened for possible confounding variables such as pathology. Incorporating pathology consistently increased the number of differentially expressed genes detected. However, the analysis was often restricted to comparisons of 1 sample per condition, which limited the power to detect changes.

Stratification is not just important in identifying more differentially expressed genes but is needed to account for confounding effects. At least one publication(81) has identified that components of the pathologies can profoundly affect the interpretation of results. For example, the presence of synovial infiltrating lymphocytes versus the specific organisation of these cells into discrete lymphoid aggregates. This has prompted the evaluation of synovial histopathology in clinical studies (73,196), and steered the design of biological drug trials including R4RA and STRAP(169,197)

This ability to retroactively stratify patients should allow further work to re-examine some of the previously published datasets outlined in Chapter 2.6. This may reveal novel pathways to target. For example, molecular pathway analysis identified Wnt signalling in diffuse synovitis as a common pathway targeted by therapy. Similar responses to therapy were also linked to the IL-2 (and presumably other IL-2-related cytokines) system and TLR signalling.

The principle aim of this chapter was to test whether inclusion of pathology would improve on the classification of differentially expressed genes following treatment. What wasn't possible with the small sample sizes was the generation of a signature that predicts response. Future work with larger cohorts, particularly those with predefined pathologies to ensure balanced datasets, may well allow the development of prognostic signatures for therapeutic response.

8. Evaluating the involvement of Jak-STAT signalling in synovitis.

8.1. Introduction.

Synovitis is associated with elevated levels of inflammatory cytokines, including various interleukins, interferons and growth factors (37,39,155,177,198–200). These cytokines activate cells through receptor systems that signal *via* a diverse array of transcription factors responsible for the control of proliferation, survival, differentiation and specific gene regulation. For example, Nuclear Factor kappa-B (NF-κB), RAR-related Orphan Receptor gamma (RORγ) and transcription factors linked to the mitogen-activated protein kinase cascade and the Janus-activated kinase–Signal Transduction and Activator of Transcription (Jak-STAT) pathway (87,149,201–204). Biological drugs and small molecule inhibitors used in the treatment of immune-mediated inflammatory diseases often inhibit these signalling events. These include the Jak inhibitors tofacitinib, baracitinib and ruxolitinib, and tocilizumab, sarilumab, siltuximab, mavrilimumab and lenzilumab, which target cytokines that signal *via* the Jak-STAT pathway (204,205).

The Jak-STAT pathway has evolved to sense and interpret cytokine cues essential for tissue and immune homeostasis. In RA, cytokines acting *via* the Jak-STAT pathway promote autoimmunity and tissue inflammation and are targeted by biological drugs (e.g., tocilizumab) or oral inhibitors (e.g., tofacitinib) prescribed in clinical practice. As part of their mode-of-action, these drugs block cytokine signalling through STAT1 and STAT3 transcription factors. In murine models of RA, these transcription factors contribute to the control of leukocyte recruitment, synovial hyperplasia, joint erosion, and T-cell driven autoimmunity(202,206–211). Both STAT1 and STAT3 have been demonstrated to have increased levels of active phosphorylation in the inflamed synovium of RA patients(203,206,207,212,213). However, given the complex nature of synovitis seen in humans it is currently unclear how these transcriptional mechanisms shape the course of inflammation to drive disease heterogeneity.

STAT1 and STAT3 transcription factors often display opposing actions in immune and stromal cells(214,215). Here, genetic ablation studies show that STAT1 and STAT3 share a complex working relationship and often oppose one another[19,37-39]. In this regard, STAT1 activities are often protective. For example, reducing proliferation and inducing apoptosis of recruited inflammatory cells(213). In contrast, STAT3 promotes inflammation through the production and secretion of pro-inflammatory cytokines and other mediators, the initiation of cellular hyperplasia and resistance to apoptosis(202,216).

To identify the contribution of the Jak-STAT pathway in each of the datasets investigated in this thesis, it is important to have an accurate list of STAT-associated genes. However, bioinformatic resources from pathway tools(217–219), as well as studies utilising ChIP-seq(220–223) find significant differences in the numbers of genes associated, with limited overlapping genes. Thus, investigations presented in the Chapter aimed to generate a bespoke list of STAT1 and STAT3 target genes that could then be applied to human transcriptomic datasets to document the involvement of these transcription factors in each synovial pathotype.

8.2. Hypothesis.

It is hypothesized that the Jak-STAT pathway plays an integral role in the development of synovitis and steers the expression of discrete gene signatures that underpin the development of pauci immune, diffuse or follicular synovitis.

8.3. Aims.

Experiments outlined in this Chapter aimed to identify the STAT regulated gene signatures linked to each of the synovial pathologies seen in RA. Further analysis evaluated the expression of these genes in specific cell populations contributing to synovitis.

8.4. Materials and methods

8.4.1. STAT1 and STAT3 associated genes from pathway analysis datasets.

Pathway analysis tools provide curated lists of genes associated with STAT1 or STAT3 activity. Gene lists were extracted from ingenuity pathway analysis (IPA)(217), Panther(219), and the Harmonizome(218) databases.

Table 8.1: ChIP-seq experiments for STAT1 or STAT3 were identified from the ENCODE encyclopaedia and literature.

Accession identifies either the ENCODE experiment or the paper from which the data was obtained. Species, cell type, and stimulation were derived from data associated with the experiment, the number of peaks mapped to genes either used published data or were mapped using the Peak Annotation and VISualisation (PAVIS) software using default parameters.

	Accession	Species	Cell type	Stimulation	Number of peaks mapped to gene
STAT1	ENCSR000DZM	Human	GM12878	NA	5983
	ENCSR000EZK	Human	HeLa-S3	IFNγ for 30min	5966
	doi: 10.4137/GRSB.S11433	Human	HeLa-S3	IFNγ for 30min	1441
	ENCSR000FAV	Human	K562	IFNα for 30min	795
	ENCSR000FAU	Human	K562	IFNα for 6H	720
	ENCSR000EHK	Human	K562	IFNγ for 30min	533
	ENCSR000EHJ	Human	K562	IFNγ for 6H	1706
	doi: 10.1038/s41590-019-0350-0	Mouse	Tnv	NA	78
	doi: 10.1038/s41590-019-0350-0	Mouse	Tnv	IL-6 for 30min with CD3/CD28	361
	doi: 10.1038/s41590-019-0350-0	Mouse	Tem	IL-6 for 30min with CD3/CD28	46
STAT3	ENCSR000DOZ	Human	MCF 10A	0.01% ethanol	5695
	ENCSR000DOQ	Human	MCF 10A	4-hydroxy-tamoxifen for 12 hours	11076
	ENCSR000DPB	Human	MCF 10A	4-hydroxy-tamoxifen for 36 hours	12707
	ENCSR000DZV	Human	GM12878	NA	3734
	ENCSR000EDC	Human	HeLA-S3	NA	2930
	doi: 10.1016/j.immuni.2010.05.003	Mouse	Tcell	CD3, CD28, IL-6, TGFβ, IFNγ for 72H + IL6 restimulation for 1H	3176
	doi: 10.1182/blood-2011-09-381483	Mouse	Macrophage	IL-10 for 4H	1103
	doi: 10.1016/j.cell.2008.04.043	Mouse	ESC	LIF and BMP4	1156
	doi: 10.1038/s41590-019-0350-0	Mouse	Tnv	NA	16
	doi: 10.1038/s41590-019-0350-0	Mouse	Tnv	IL-6 for 30min with CD3/CD28	218
	doi: 10.1038/s41590-019-0350-0	Mouse	Tem	IL-6 for 30min with CD3/CD28	84

8.4.2. *Identification of STAT1 and STAT3 gene targets from publicly available ChIP-seq results.*

Chromatin-immunoprecipitation sequencing (ChIP-seq) datasets for STAT1 and STAT3 were obtained from the ENCODE encyclopaedia(220). Additional results were obtained by searching the literature for STAT1 or STAT3 ChIP-seq experiments. To obtain a list of associated genes, peaks from the ENCODE experiments were uploaded to the Peak Annotation and Visualisation (PAVIS) tool(129) using the default parameters of 5kb upstream and 1kb downstream of the Transcription Start Site (TSS) to identify all peaks that fell into these regions. For literature results, the associated genes were obtained from supplemental data available with each publication. A summary of this information is presented in Table 8.1, listing the cell type, and stimulation, as well as the number of peaks mapped to specific gene loci.

8.4.3. Tracking STAT1 and STAT3 involvement in synovitis.

To investigate the role of STAT1 and STAT3 in synovial pathology, datasets within the GSE48780 cohort were interrogated using the STAT-associated gene list available through IPA. Additional analysis of the GSE45867 cohort was used to examine how biological drug targeting of the IL-6R modifies the expression of STAT regulated genes.

8.4.4. ChIP-seq analysis of STAT1 and STAT3 in murine antigen-induced arthritis.

ChiP-seq was performed on synovial tissue from mice with antigen-induced arthritis (protocols described in Chapter 2). Synovial samples were collected from naïve unchallenged mice, and at day-3 (reflecting early stage disease) and day-10 (reflecting late stage disease) of antigen-induced arthritis in wild type (WT), *Il6ra*$^{-/-}$ and *Il27ra*$^{-/-}$ mice.

8.5. Results.

8.5.1. Common STAT1 and STAT3 associated genes from pathway analysis datasets

Inspection of the computational toolkits within Harmonizome, IPA, and Pather identified a panel of genes affiliated with STAT1 and STAT3 signalling. From each of these databases a total of 643, 230, and 634 unique genes where identified for STAT1, and an additional 1366, 396, and 819 unique genes for STAT3. The Venn diagrams presented in Figure 8.1 highlight the small overlap seen between datasets. Analysis identified 2 shared genes for STAT1 (*MYC* and *PDGFRB*), and a 11 genes for STAT3 (*ANGPT2, CCL8, CDKN1A, ID2, IKBKE, KAT2B, LTA, MYC,* *PGR, RBPJ,* and *SNAI1*).

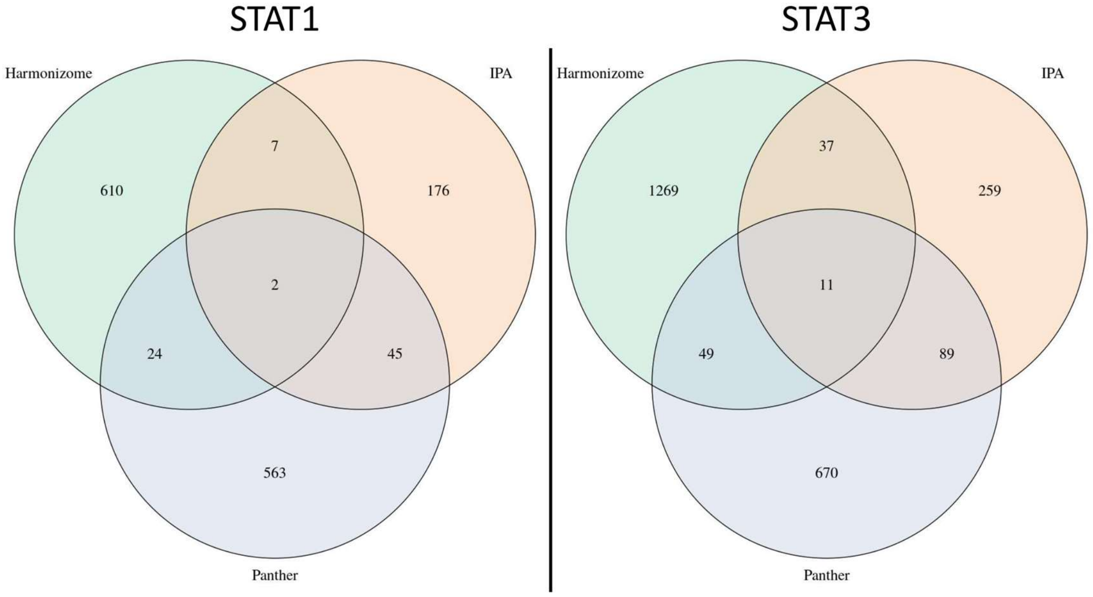


Figure 8.1: Venn diagram of STAT associated genes from pathway analysis databases.

STAT1 or STAT3 associated genes were extracted from the pathway analysis tools Harmonizome, Ingenuity Pathway Analysis (IPA) and Panther. Numbers on the Venn diagram represent the numbers of genes from each database and how they overlap

8.5.2. Identification of STAT1 and STAT3 associated genes from ChIP-seq experiments.

ChIP-seq experiments for STAT1 or STAT3 were obtained from human and mouse sources, comparison of the genes associated with all the STAT1 ChIP-seq datasets revealed no common core genes, however, when restricted to human samples only, this results in 17 core genes (*AIM2, APOL1, APOL6, BAZ2A, HLA-E, ICAM1, IRF9, ITPR1, KSR1, NAPA, NCOA7, OTOF, RUNX1, SEMA4B, SP140L, STAT1*, and *WARS*). STAT3, however, has 9 core genes that are found in both the human and mouse datasets (*ARHGEF12, BCL3, CDK6, CDKAL1, NFKBIZ, SOCS3, STAT3, ZFP36*, and *ZFP36L1*).

Table 8.2 and 8.3 show the numbers of overlapping genes for STAT1 and STAT3 respectively. Broadly, both tables make it clear that sharing stimulus results in more overlapping genes, with some associated to specific cell types.

.

Table 8.2: Number of STAT1 associated genes derived from ChIP-seq experiments and how they overlap when comparing conditions. Samples compared is illustrated by the percentage of overlaps (from sample total peaks) listed in the main table.

	Sample	GM12878	HeLa-S3	HeLa-S3	K562	K562	K562	K562	Tnv	Tnv	Tem
	Stimulus	-	IFNγ	IFNγ	IFNα	IFNα	IFNγ	IFNγ	-	IL-6	IL-6
	# genes	5983	5966	1441	795	720	533	1706	78	361	46
Comparison	# overlaps										
HELA-S3	1236		20.7%	85.8%							
IFNγ	179		3.0%	12.4%			33.6%	10.5%			
K562	59				7.4%	8.2%	11.1%	3.5%			
K562-IFNα	297				37.4%	41.3%					
K562-IFNγ	318						59.7%	18.6%			
IL-6	2									0.6%	4.3%
Epithelial	363	6.1%	6.1%	25.2%							
Human	17	0.3%	0.3%	1.2%	2.1%	2.4%	3.2%	1.0%			

Table 8.3: Number of STAT3 associated genes derived from ChIP-seq experiments and how they overlap when comparing conditions.

	Sample	GM12878	HeLA-S3	MCF 10A	MCF 10A	MCF 10A	Tcell	ESC	Tnv	Tem	Macrophage
	Stimulus	-	-	EtOH	TAM	TAM	Mix	LIF/BNP4	IL6	IL6	IL10
	# genes	3734	2930	5695	11076	12707	3176	1156	218	84	1103
Comparison	# overlaps										
Tamoxifen	10039				90.6%	79.0%					
Epithelial	476	12.7%	16.2%	8.4%	4.3%	3.7%					
IL6	14								6.4%	16.7%	
MCF 10A	5127			90.0%	46.3%	40.3%					
T-cell	10						0.3%		4.6%	11.9%	
Lymphocyte	4						0.1%		1.8%	4.8%	0.4%
Human	476	12.7%	16.2%	8.4%	4.3%	3.7%					
Mouse	54						0.1%	0.3%	1.8%	4.8%	0.4%
All	9	0.2%	0.3%	0.2%	0.1%	0.1%	0.3%	0.8%	4.1%	10.7%	0.8%

8.5.3. Identification of STAT1 and STAT3 involvements in synovial pathology.

8.5.3.1. GSE48780 – Stratified established disease.

Adopting the gene signatures associated with STAT1 and STAT3 activity, computational analysis evaluated the expression of these genes in each of the synovial pathologies. This approach identified 3 well defined clusters of gene expression. As represented in Figure 8.2, genes affiliated with Cluster 1 showed considerable overlap with both diffuse and pauci-immune synovitis. In contrast, Cluster 2 was predominantly linked with diffuse and follicular synovitis and Cluster 3 was more closely associated with follicular pathology. Thus, Jak-STAT signalling through STAT1 and STAT3 transcription factors appear to be intrinsically linked with each form of synovitis and control unique sets of genes involved in pauci immune, diffuse and follicular synovitis. In this regard, many of the genes linked to these individual clusters identify with factors that shape the course of disease. For example, analysis of Cluster-1 identified 202 genes (388 probe-sets), which are affiliated with VEGFA-VEGFR2 signalling, the control of focal adhesion through the PI3K-Act-mTOR and AGE-RAGE pathways and the TGFß regulation of the extracellular matrix. Genes seen in Cluster-2 comprised 240 genes (453 probe-sets) linked with type-II interferon signalling, whereas Cluster 3 had 170 genes (264 probe-sets) associated with the control of lymphocyte survival, the differentiation of T-cells towards a Th17 lineage and the regulation of lymphokine activities.

8.5.3.2. GSE45867 – Before and after tocilizumab treatment.

With the limited number of samples, there is no clear difference between follicular and diffuse pathologies expression of STAT associated genes (Figure 8.3). Pauci-immune samples do demonstrate lower STAT associated gene expression overall, which is most clearly demonstrated in Cluster-2. Cluster-1 does not demonstrate any clear pattern of expression when considering pathology or response to tocilizumab. Cluster-2 has 282 genes (524 probes) that show strong expression in the follicular and diffuse pathologies, and these are “switched off” following tocilizumab treatment. Following treatment, these follicular and diffuse patients now resemble those of pauci-immune after a successful response to therapeutics. The one patient who failed to respond to therapy showed a maintained expression of STAT associated genes in Cluster-2. Pathway analysis of Cluster-2 indicates it is associated with T-cell receptor regulation of apoptosis, and type II interferon signalling.

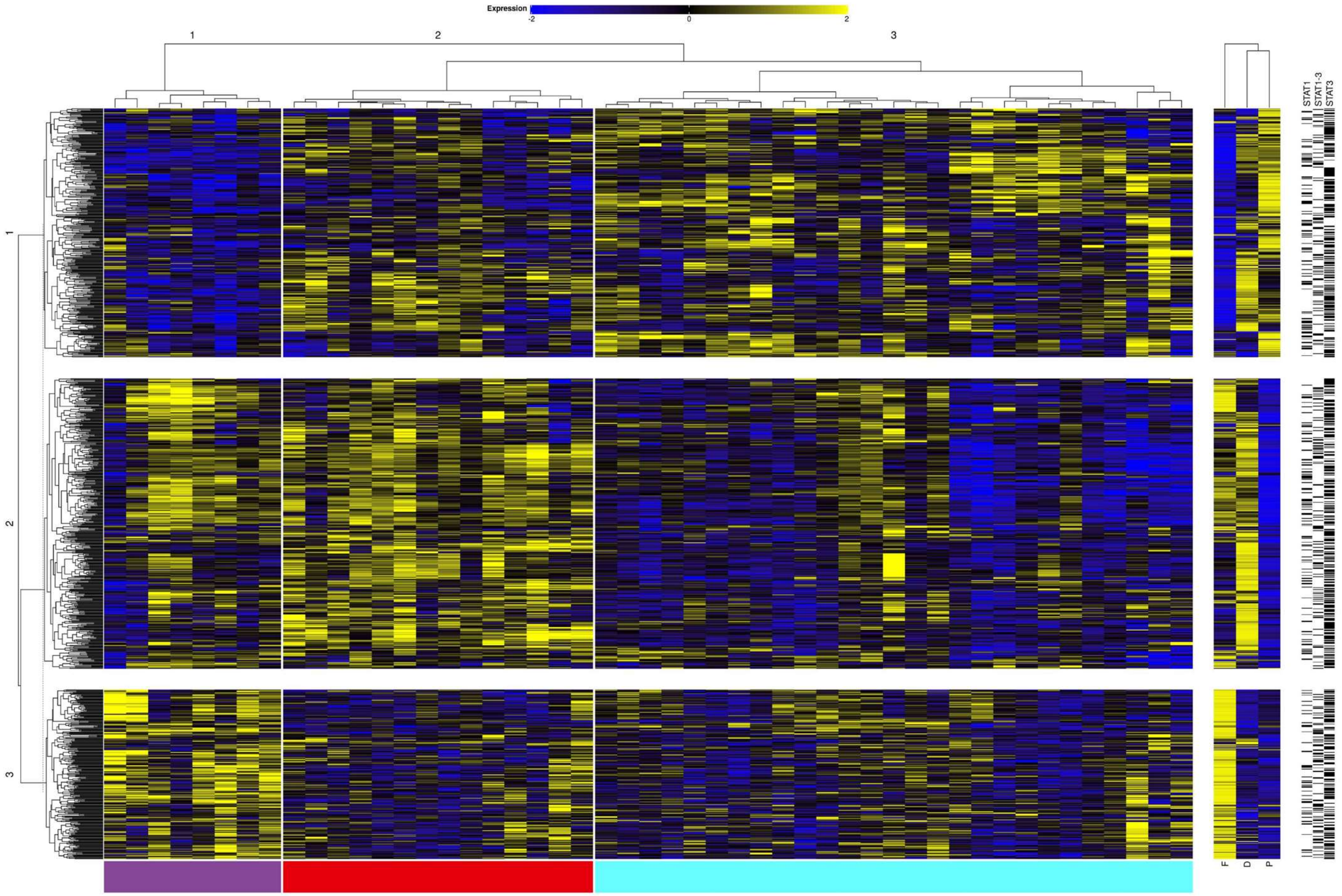


Figure 8.2: Expression of STAT1 and STAT3 associated genes in stratified established disease.

Clustering the associated probes identifies three clusters of genes that are associated with the three pathologies of rheumatoid arthritis. Cluster 1 has mixed expression in diffuse (red) and pauci-immune (cyan) samples, which are associated with pathways linked to stromal growth and modification of the extracellular matrix. Cluster2 is primarily associated with the diffuse with some overlap into follicular (purple) pathologies, and is heavily enriched for type II interferon signatures. Cluster 3 is strongly expressed in follicular and linked to control of apoptosis and T-cell differentiation.

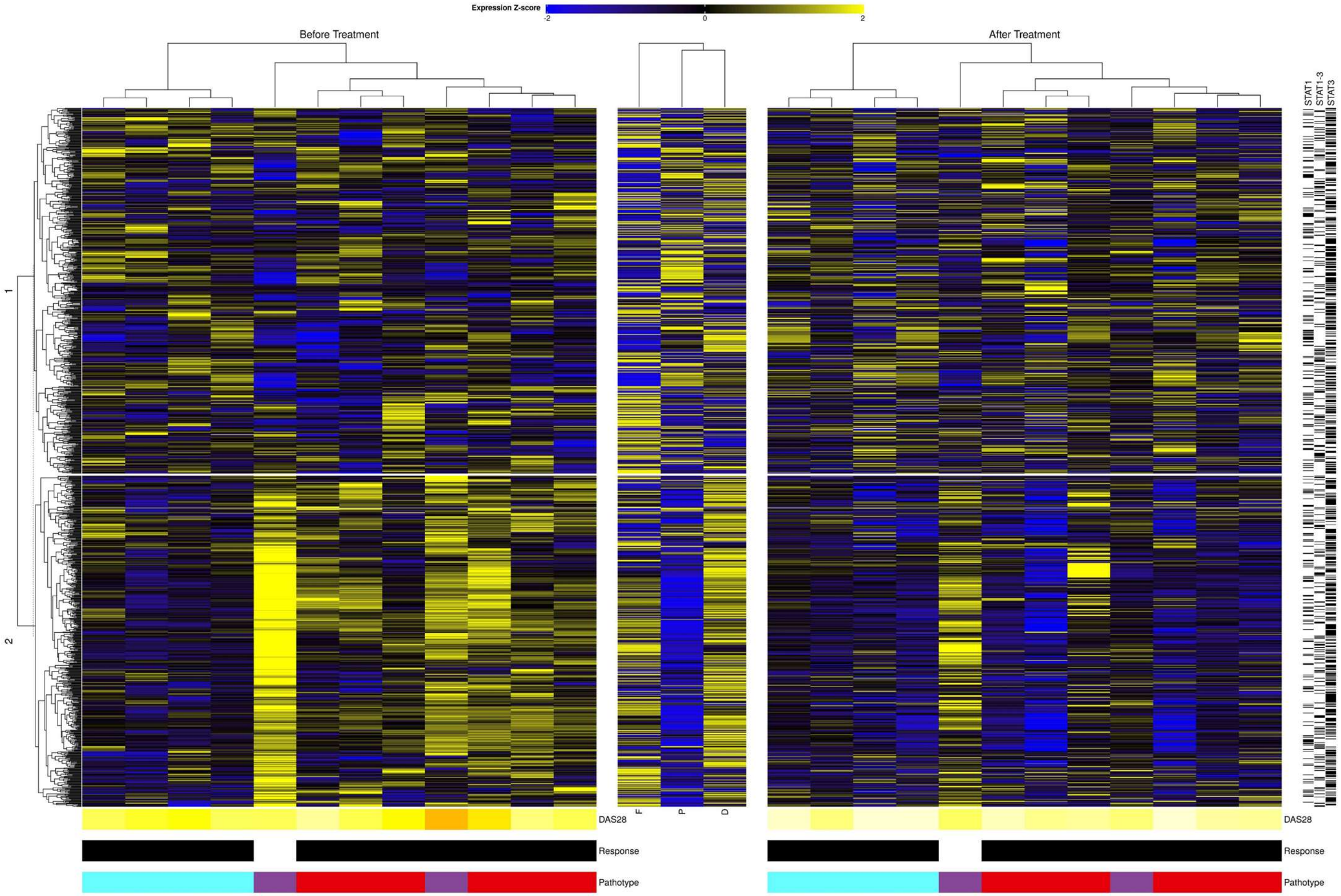


Figure 8.3: Expression of STAT1 and STAT3 associated genes before and after tocilizumab treatment.

Before treatment, the pathologies demonstrate differential behaviour in cluster 2, with follicular (purple) and diffuse (red) showing high expression and pauci-immune (cyan) having low expression. Following treatment responders in the follicular and diffuse patients now have a pauci –immune like profile. These STAT associated genes in cluster 2 are associated with T-cell receptor signalling in apoptosis and type II interferon signalling.

8.5.4. STAT1 and STAT3 genes from ChIP-seq of murine AIA.

Given the discrepancy in STAT associated genes covered in sections 8.5.1 and 8.5.2, it becomes apparent that to understand the role of STAT1 and STAT3 in the pathogenesis of RA, it is essential to identify the STAT associated genes from inflamed synovial tissue. Therefore, libraries were prepared from STAT1 or STAT3 immunoprecipitated DNA derived from the inflamed synovial tissue of mice with AIA. Library quality was assessed using bioanalyser (Figure 8.4) and nanodrop and was deemed sufficient to proceed to sequencing.

Table 8.4: Mapping statistics for the STAT1 and STAT3 immunoprecipitations

	Number of read pairs	After trimming (%)	After mapping (%)	Forward Strand (%)	Reverse strand (%)	Duplicated (%)
WT A I	78615864	77.7	77.2	50.2	49.8	79.9
WT A S1	84125691	92.3	92.0	50.2	49.8	72.6
WT A S3	51215919	85.2	84.7	50.3	49.7	65.8
Il-6r$^{-/-}$ A I	84664897	94.5	93.7	50.4	49.6	73.4
Il-6r$^{-/-}$ A S1	51123107	77.9	77.4	50.2	49.8	33.1
Il-6r$^{-/-}$ A S3	72830052	86.5	86.0	50.2	49.8	69.1

Table 8.4 contains the alignment statistics, which demonstrates the good mapping of the trimmed reads, with no bias towards the forward or reverse strand. However, a large percentage of duplicates (30-70%) were identified within the datasets. Figure 8.5 illustrates the fastqc report using one of the samples (*Il6ra*$^{-/-}$ STAT1), which shows excellent quality reads across the length and full-length 75bp reads, but a high percentage of GC content that doesn't match the theoretical distribution. Therefore, sequencing accuracy was high for the reads, but the DNA utilised for the library preparation was not very diverse (high duplicates).

Table 8.5: Number of peaks detected by max at each threshold value.

	Number of Peaks		
	Q = 0.1	Q = 0.05	Q = 0.01
WT A S1	1504	724	431
WT A S3	172	167	127
Il6ra$^{-/-}$ A S1	218	139	109
Il6ra$^{-/-}$ A S3	209	184	75

After mapping, peaks were called on the samples with duplicates removed, with the number of peaks detected outlined in Table 8.5 for each of the thresholds. The association between the immunoprecipitated peaks and genes was identified using PAVIS, to maximise the number of associated genes the threshold value of q=0.1 was utilised, the number of associated genes is listed in Table 8.6. Visualising these associated genes demonstrates that only 20-30% of peaks are associated with genes, as shown by that large percentage of non-associated peaks (dark grey) in Figure 8.6. Unsurprisingly, most of the peaks (70-80%) associated with genes are protein-coding, but there are also a considerable percentage of sequencing peaks affiliated with long intergenic non-coding RNA (12-20%).

Table 8.6: Number of associated genes identified by PAVIS using default parameters of 5kb upstream and 1kb downstream using MACS Q=0.1 results

	Linked genes
WT A S1	321
WT A S3	25
Il-6r$^{-/-}$ A S1	31
Il-6r$^{-/-}$ A S3	36

Investigating some of these peaks directly in Integrative Genomics Viewer (IGV)(224) does highlight some good peaks, an example of which is shown in Figure 8.7. Unfortunately, this peak is not associated with any genes, but it does demonstrate a good enrichment of reads in the immunoprecipitated sample (pink or green) compared to the input (grey) when looking at the BAM coverage. The numbers in square brackets on the plots illustrate the highest depth of reads seen at these genomic loci.

However, exploring further into the dataset reveals some issues with the data, most of the peaks are associated with low numbers of reads 10-20 reads that are difficult to differentiate from general noise. However, there are several locations in the genome that have excessive numbers of reads, not just in the immunoprecipitated samples but also the input sample. An example of this is shown in Figure 8.8, where high proportion of the total reads (1-2%) are observed in a single small region (10kb), this was associated with all conditions. Where peaks elsewhere have reads that vary from 10 to 150 read depths, the peaks in Figure 8.8 have a peak read depth of 17.5 thousand reads, which is considerably higher than average. Moreover, these peaks exhibit the same pattern of mapping in all of the samples - from both genotypes (WT, *Il6ra*$^{-/-}$) and for input and immunoprecipitated DNA. This clearly indicates some kind of technical bias in the sample preparation that has not been determined.

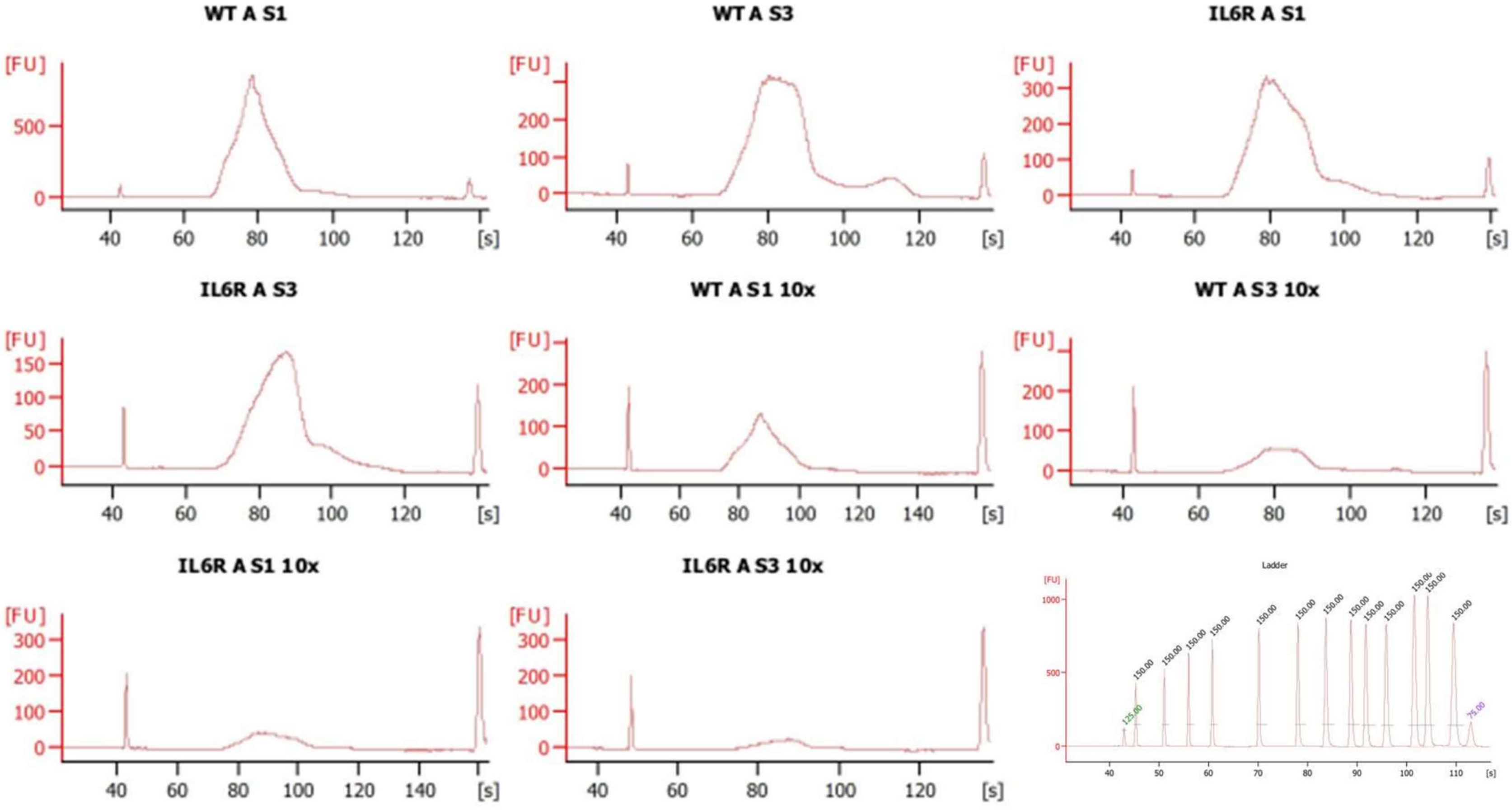


Figure 8.4: Bioanalyser report on library quality of immunoprecipitated samples.

Peaks in samples fell between 70 and 90 seconds which translates to between 300 and 700 base pairs.

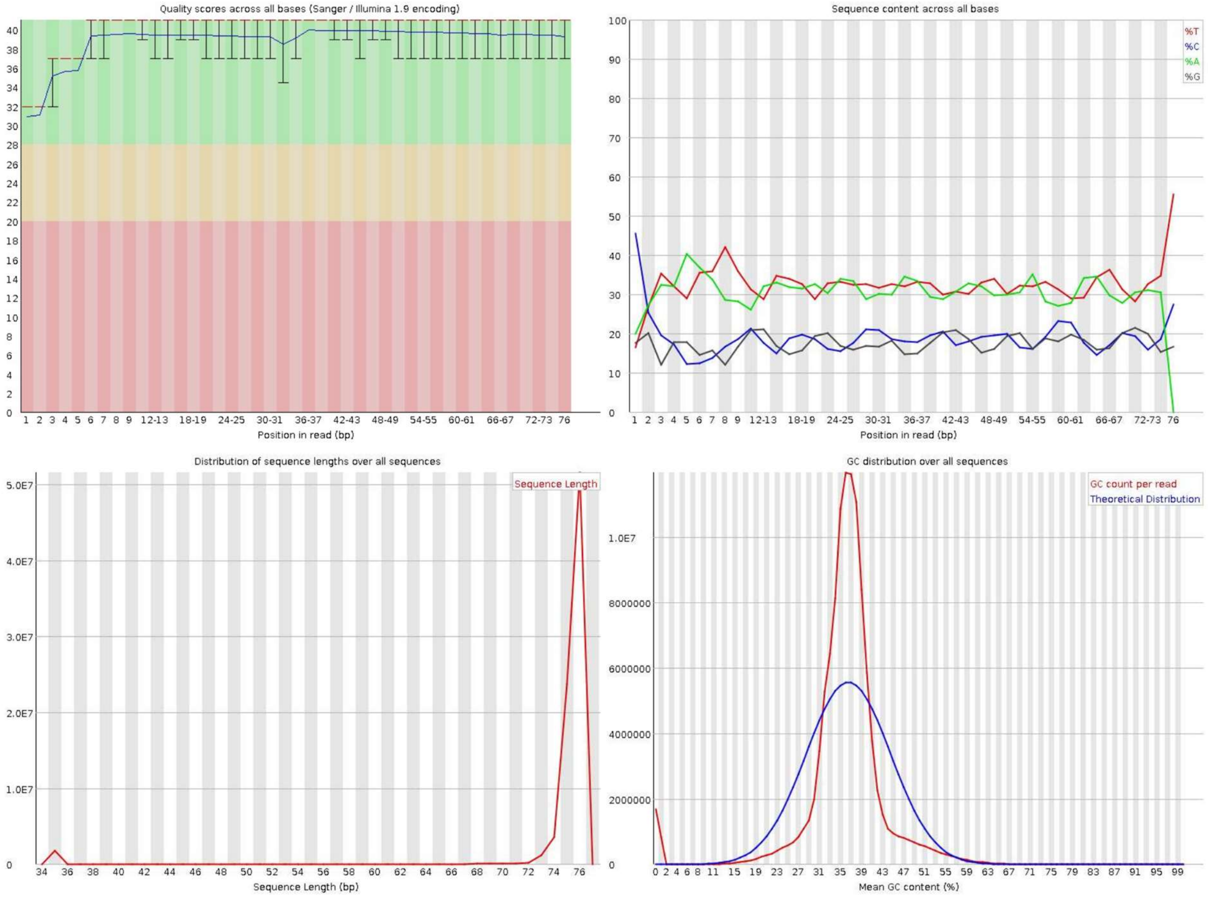


Figure 8.5: Typical Fastqc results for ChIP-seq run.

In this case for the Il6r-/- A STAT1 sample. Reads showed high quality across the whole length with phred scores above 30 for entire reads(A), which all maximised the 75bp sequencing length(B). Per base content isn't ideal(C), and this is reflected in the deviation from the theoretical distribution of GC content(D).

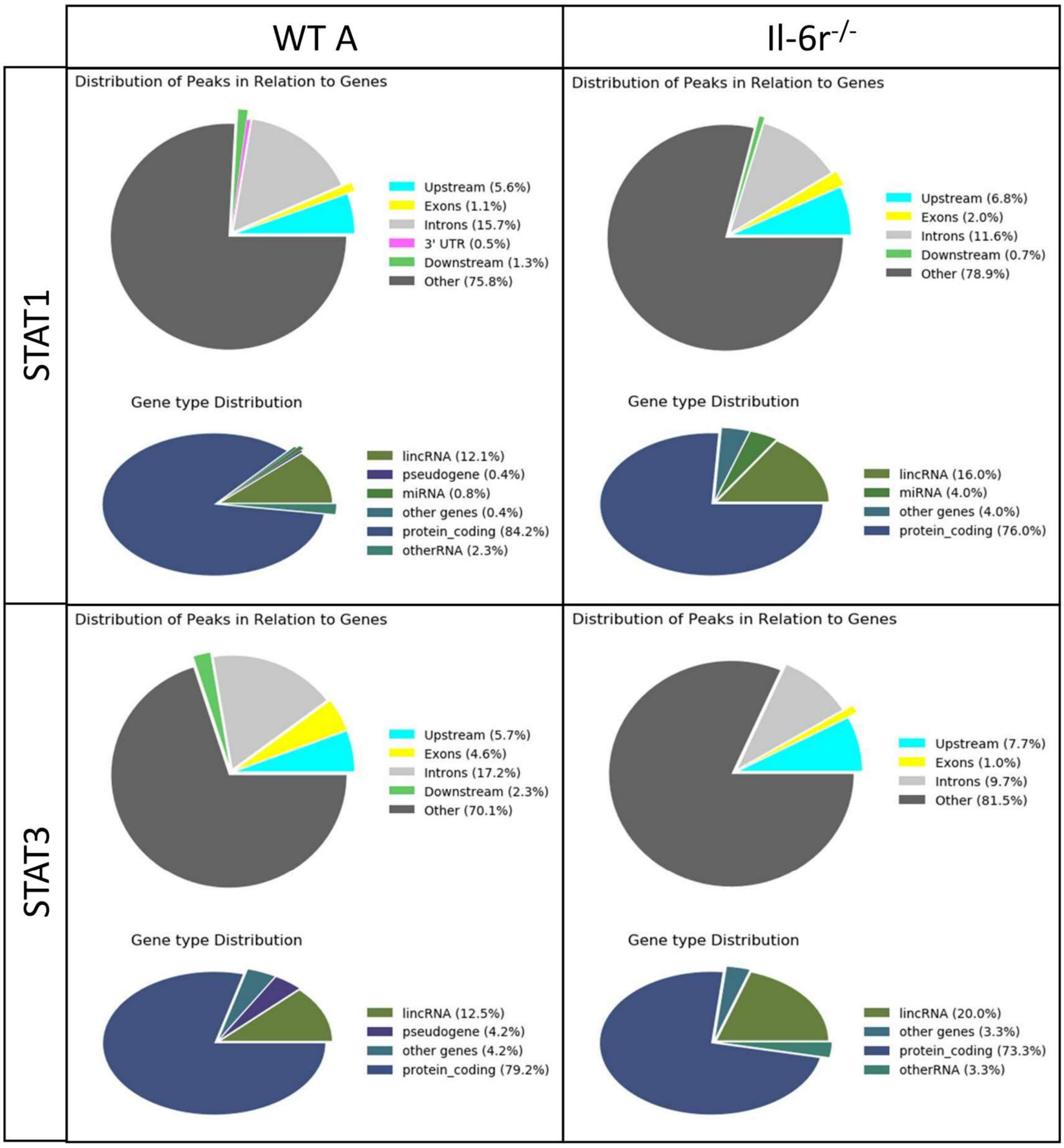


Figure 8.6: Summary of the association between immunoprecipitated peaks and genes.

The first pie chart illustrates the distribution of peaks associated with genes – using the default parameters of PAVIS. For all samples the majority of peaks are not associated with genes (shown in dark grey). Of those peaks that overlapped with genes, the majority were associated with protein coding genes (dark blue), but also a surprising number of long intergenic non-coding RNA.

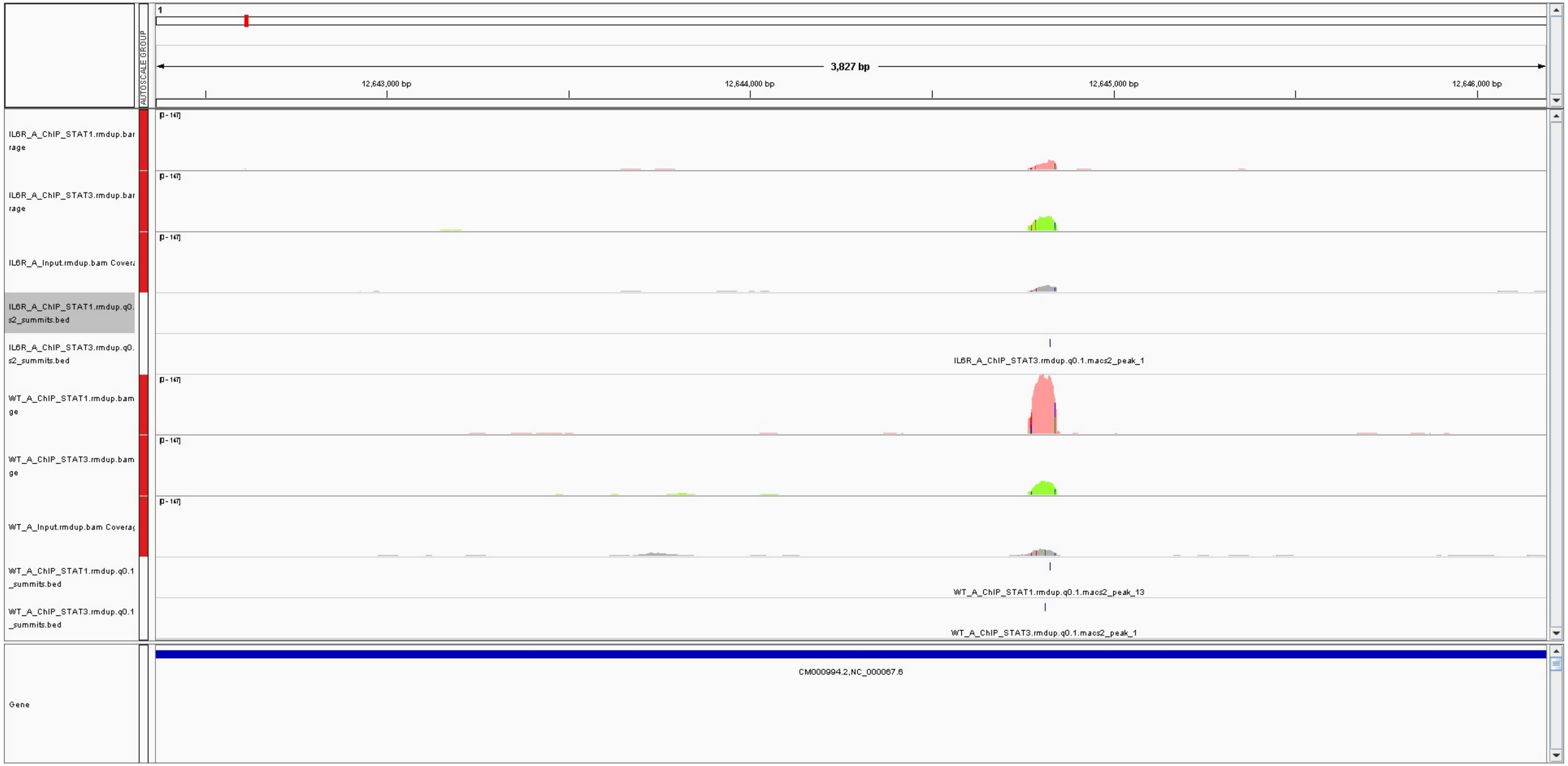

1
AUTOSCALE GROUP
3,827 bp
12,643,000 bp
12,644,000 bp
12,645,000 bp
12,646,000 bp
IL6R_A_ChIP_STAT1.rmdup.bar rage
[0 - 147]
IL6R_A_ChIP_STAT3.rmdup.bar rage
[0 - 147]
IL6R_A_Input.rmdup.bam Cover
[0 - 147]
IL6R_A_ChIP_STAT1.rmdup.q0. s2_summits.bed
IL6R_A_ChIP_STAT3.rmdup.q0. s2_summits.bed
IL6R_A_ChIP_STAT3.rmdup.q0.1.macs2_peak_1
WT_A_ChIP_STAT1.rmdup.bam ge
[0 - 147]
WT_A_ChIP_STAT3.rmdup.bam ge
[0 - 147]
WT_A_Input.rmdup.bam Coverag
[0 - 147]
WT_A_ChIP_STAT1.rmdup.q0.1 _summits.bed
WT_A_ChIP_STAT1.rmdup.q0.1.macs2_peak_13
WT_A_ChIP_STAT3.rmdup.q0.1 _summits.bed
WT_A_ChIP_STAT3.rmdup.q0.1.macs2_peak_1
CM000994.2,NC_000067.6
Gene

Figure 8.7: An example of a good ChIP-seq peak.

WT STAT1 exhibits a large depth of reads that clearly exceeds that of the input sample and is therefore called as a peak, peaks are also called in all samples, though WT STAT1 is the clearest. Unfortunately, this peak is not associated with any genes

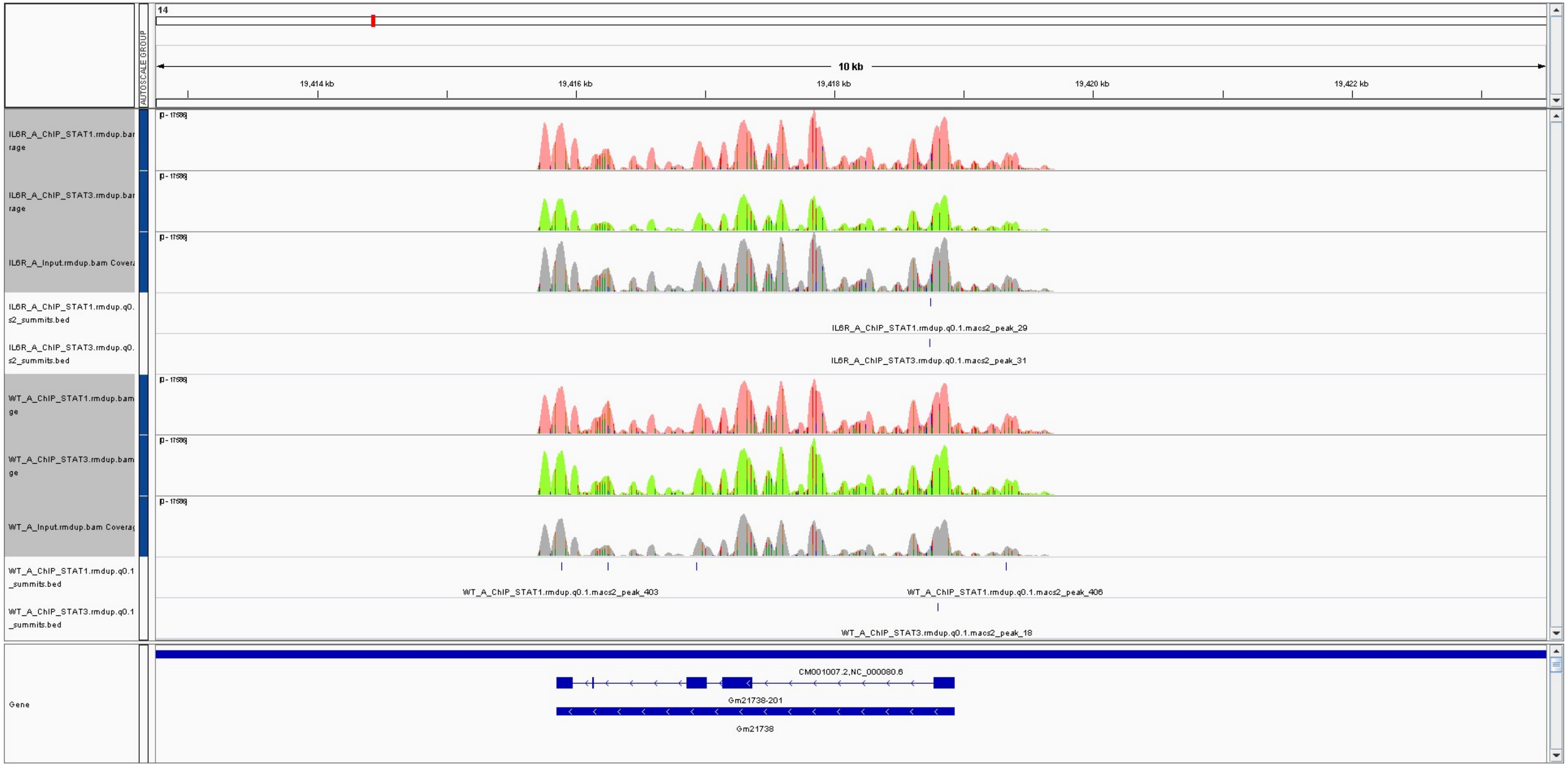

14
AUTOSCALE GROUP
10 kb
19,414 kb
19,416 kb
19,418 kb
19,420 kb
19,422 kb
IL6R_A_ChIP_STAT1.rmdup.bar rage
IL6R_A_ChIP_STAT3.rmdup.bar rage
IL6R_A_Input.rmdup.bam Covera
IL6R_A_ChIP_STAT1.rmdup.q0. s2_summits.bed
IL6R_A_ChIP_STAT3.rmdup.q0. s2_summits.bed
WT_A_ChIP_STAT1.rmdup.bam ge
WT_A_ChIP_STAT3.rmdup.bam ge
WT_A_Input.rmdup.bam Coverag
WT_A_ChIP_STAT1.rmdup.q0.1 _summits.bed
WT_A_ChIP_STAT3.rmdup.q0.1 _summits.bed
Gene
[0 - 17586]
IL6R_A_ChIP_STAT1.rmdup.q0.1.macs2_peak_29
IL6R_A_ChIP_STAT3.rmdup.q0.1.macs2_peak_31
WT_A_ChIP_STAT1.rmdup.q0.1.macs2_peak_403
WT_A_ChIP_STAT1.rmdup.q0.1.macs2_peak_406
WT_A_ChIP_STAT3.rmdup.q0.1.macs2_peak_18
CM001007.2,NC_000080.6
Gm21738-201
Gm21738

Figure 8.8: An example of the issues associated with the ChIP-seq results.

A considerable proportion (1-2%) of the total reads are found enriched in this single region – even in the input which should show no bias towards reads.

8.6. Discussion.

This chapter demonstrates that very few genes can be definitively linked to either STAT1 or STAT3, comparison of STAT associated genes from pathway tools and ChIP-seq experiments failed to provide identify a core set of genes that was consistent across all of these resources.

In looking for a core set of STAT-associated genes, the ChIP-seq results demonstrate that stimulus drives more consistent gene sets than cell origin. For example, a larger percentage of genes are associated with interferon response vs genes common to K562 cells, whilst Tamoxifen has a larger component of genes compared to the MCF 10A cell line. This reinforces the point that to understand the roles of STAT1 and STAT3 in the pathogenesis of rheumatoid arthritis, we need a signature that is reflective of the stimulus these cells received in the inflamed joint.

Moreover, these ChIP-seq results highlight the potential for batch effects to alter what genes are associated with the STATs. For example, STAT1 in HeLa-S3 cells stimulated with IFNγ has a considerable difference in the number of associated genes detected (5996 vs 1441 genes), which could be attributed to several factors such as antibodies used, peak-calling parameters etc. although these samples had a high degree of overlap for the smaller dataset (85%) and a good proportion of the larger (21%).

That said, even without a true core STAT-associated gene list, it is still possible to explore the role of STAT1 or STAT3 in synovitis. Exploring STAT-associated genes obtained from the IPA database in defined pathologies highlights distinct clusters of expression that is linked to the pathology, which is associated with numerous pathways known to be important in rheumatoid arthritis. Similarly, we see that these STAT-associated genes are “switched off” in patients who respond to treatment.

Unfortunately, ChIP-seq is somewhat of an art and ultimately demonstrates no easy positive controls. Adding to the difficulties is the AIA model takes a month to prepare and necessitates pooling of mice to provide enough tissue to analyse. Whilst the initial sequencing run identified many peaks, it also highlighted several issues with the sample preparation, resulting in an early cessation of the experiment, and overall a failure to correctly identify a STAT1 and STAT3 profile in synovitis. Ultimately, further work will be required to re-examine the role of STAT1 and STAT3 in the synovium.

9. General Discussion.

9.1. Introduction.

Rheumatoid arthritis is a heterogeneous disease with considerable variability in clinical presentation, disease progression, and associated comorbidities. As consequence, patients often display varying efficacies to treatment. Here, differences in therapeutic response to standard DMARDs or biological drugs often reflect differences in synovitis. Small needle biopsy sampling of inflamed joints has identified three discrete forms in synovial pathology – termed follicular, diffuse, and pauci-immune synovitis. Research conducted in this thesis developed several bioinformatic tools to interrogate transcriptomic data deposited in open access repositories. My ambition was to establish methodologies that could be used across different patient cohorts to improve the classification of synovitis, identify the inflammatory pathways that drive disease heterogeneity, and support predictions of clinical response to biological therapy.

9.1. Gene signatures that stratify pathology.

Several gene signatures have been discussed in this thesis, that demonstrated good performance and stratified both the follicular and pauci-immune pathologies. One of the biggest challenges for any prognostic gene signature is its validation in independent cohorts. In this thesis, I demonstrated that two signatures (17 gene, and RandomForest; Chapter 4) have performed consistently well. For example, these methods stratified follicular and diffuse synovitis across multiple independent clinical cohorts, and in both early and established disease. Figure 9.1 shows the performance of these models in each of the four independent cohorts, with the 17 gene signature exhibiting the most consistency across the datasets.

Unfortunately, the models perform poorly at classifying the diffuse pathology. Given the differences observed in the datasets as to expression of immune markers, it is quite possible that the gene signatures generated in Chapter 4 are overfitted to this dataset. Therefore, with a better-defined diffuse pathology, it may be possible to develop a classifier that works on these datasets and any future datasets released.

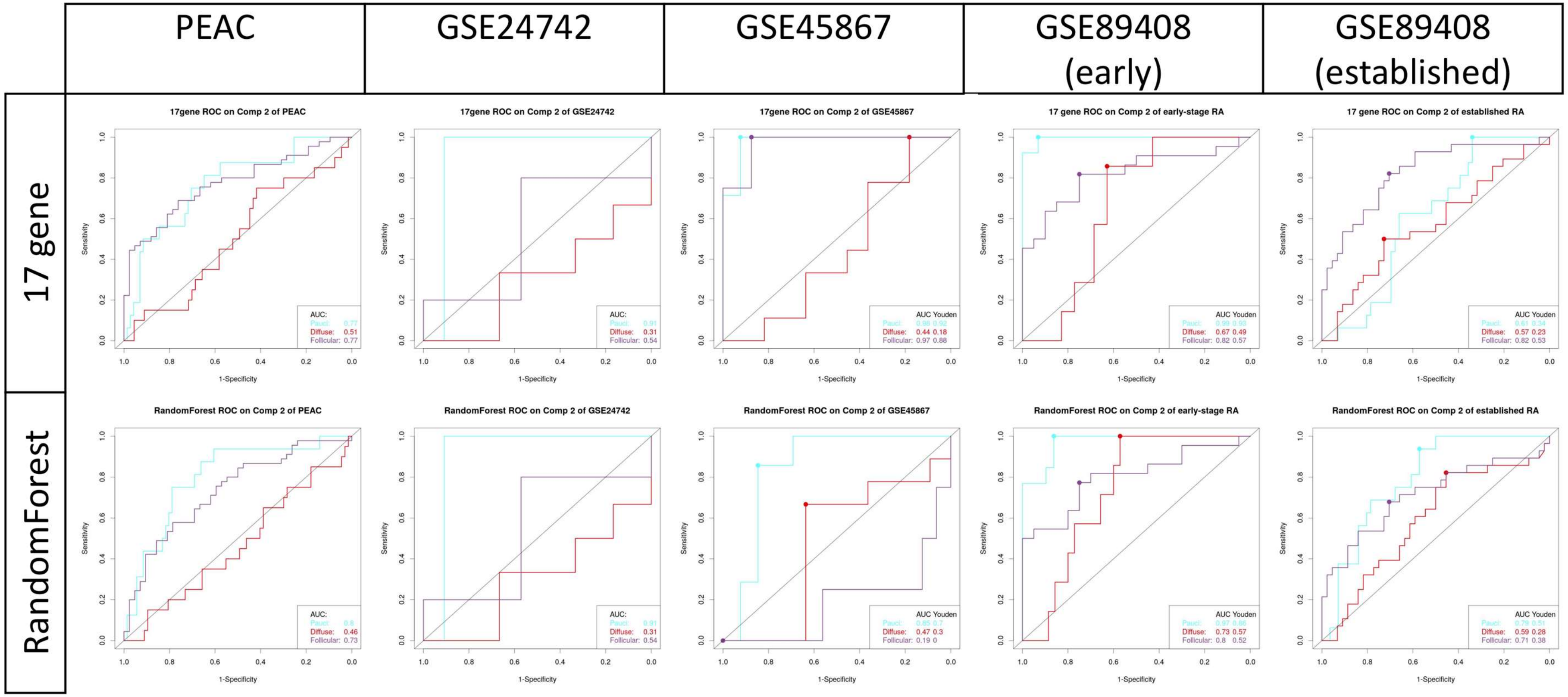


Figure 9.1: The two best models across all 4 independent datasets utilised in this thesis.

Testing the 17 gene and RandomForest classifiers across the 4 independent clinical cohorts. These samples represent both early (PEAC and GSE89408) and established disease (GSE24742, GSE45867, GSE89408). These classifiers performed consistently well at discriminating Follicular and Pauci-immune pathologies (with the exception of GSE45867 for RandomForest)

9.2. Rheumatoid arthritis pathologies have distinct transcriptional profiles.

Previous studies have demonstrated that rheumatoid arthritis is highly heterogeneous with synovial pathologies displaying alternate transcriptional profiles(73,130). Extending these investigations, I now present several key pieces of new information, which may help understand the course of synovitis development in rheumatoid arthritis.

Histological studies and interpretations from transcriptomic datasets suggest that specific stromal and immune cells play central roles in the development of follicular (lymphoid-rich), diffuse (myeloid-rich), and pauci-immune (fibroblast-rich) synovitis. However, current evaluations of synovitis have failed to establish whether:

(A) The described synovial pathologies arise through specific inflammatory pathway that drive the course of the disease

or

(B) That differences in the clinical presentation of synovitis represent individual stages within the natural trajectory of the disease.

Analysis of transcriptomic data for follicular and pauci-immune synovitis showed that the transcriptional profiles for both pathologies where highly stable and remained comparable between early and established disease. Here, interrogation of all the datasets utilised in this thesis revealed that lymphoid markers are consistently upregulated in all samples classified as follicular synovitis (see Figures 9.2 and 9.3). For example, the archetype for follicular synovitis derived from GSE48780 showed a strong expression of lymphoid lineage markers, as well a considerable component of myeloid lineage markers. This transcriptional stability is also observed in the pauci-immune samples, which generally shows low expression of immune markers across the cohorts, with a small subset of markers consistently expressed in biopsies from early and established disease (Figures 9.4 and 9.5). Thus, follicular and pauci-immune synovitis may originate from specific inflammatory pathways that steer lymphoid or fibroblast involvements

The diffuse pathology however demonstrates considerable overlap between the follicular and pauci-immune profiles (Figure 9.6 and 9.7). This fits with a more updated description of diffuse synovitis, which subsets this form of disease into “lymphoid-myeloid” and “fibroblast-myeloid” synovitis (71,130). In the dataset used to develop my bioinformatic classifiers (GSE48780) the diffuse samples fall into 2 primary clusters of probes that have limited overlap with the other pathologies. These profiles were not, however, detected in any of the other cohorts analysis (see Figure 9.7). Here, analysis of the RNA-seq datasets available in GSE89408 identified patients with diffuse synovitis with similarities to either pauci immune or follicular pathology.

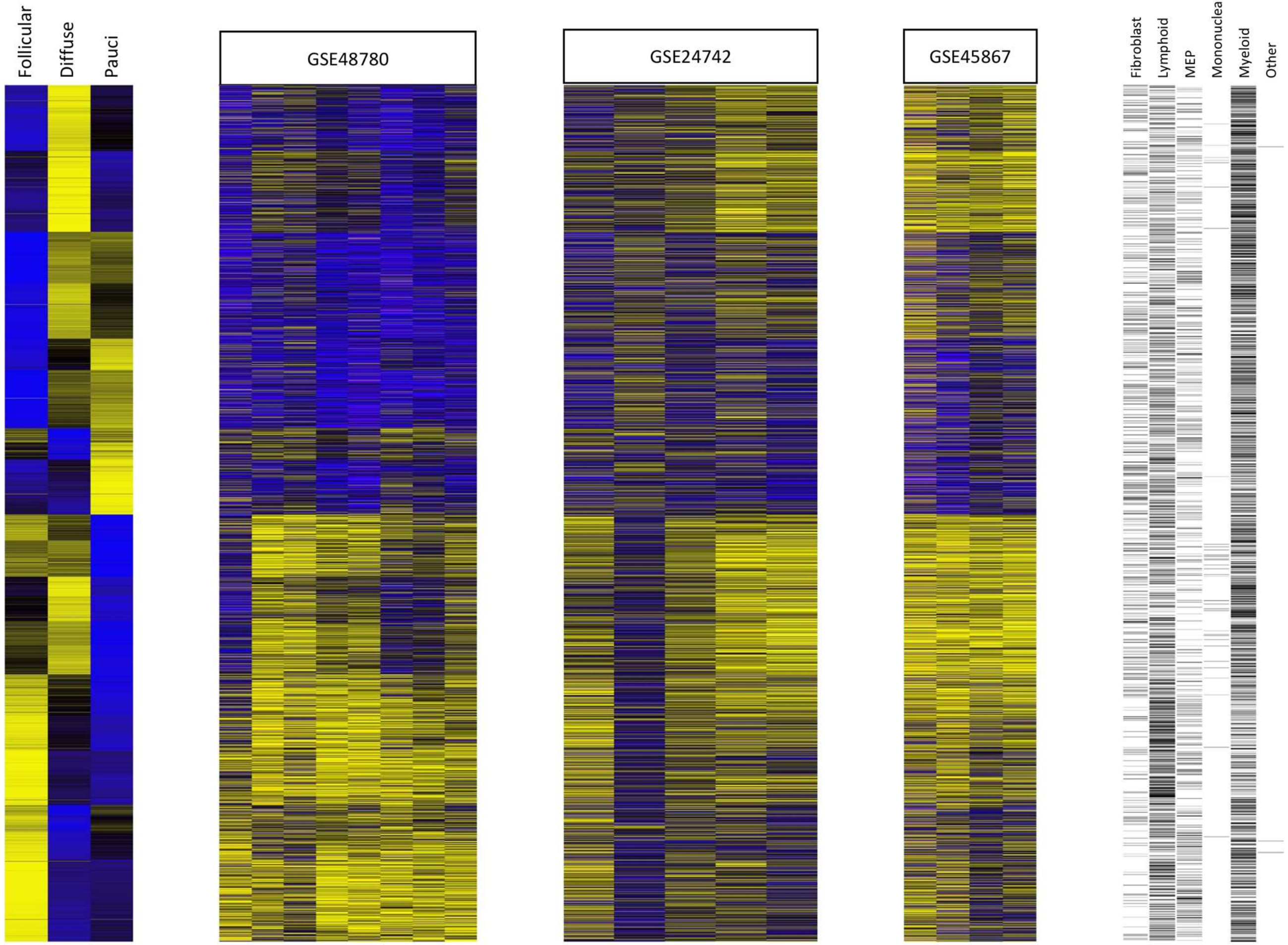


Figure 9.2: Follicular samples from the microarray based datasets GSE48780, GSE24742, and GSE45867.

Samples demonstrate consistent expression of lymphoid markers identified in the training dataset (GSE48780) but also show some overlap with the diffuse archetype in the other two datasets (GSE24742, GSE45867).

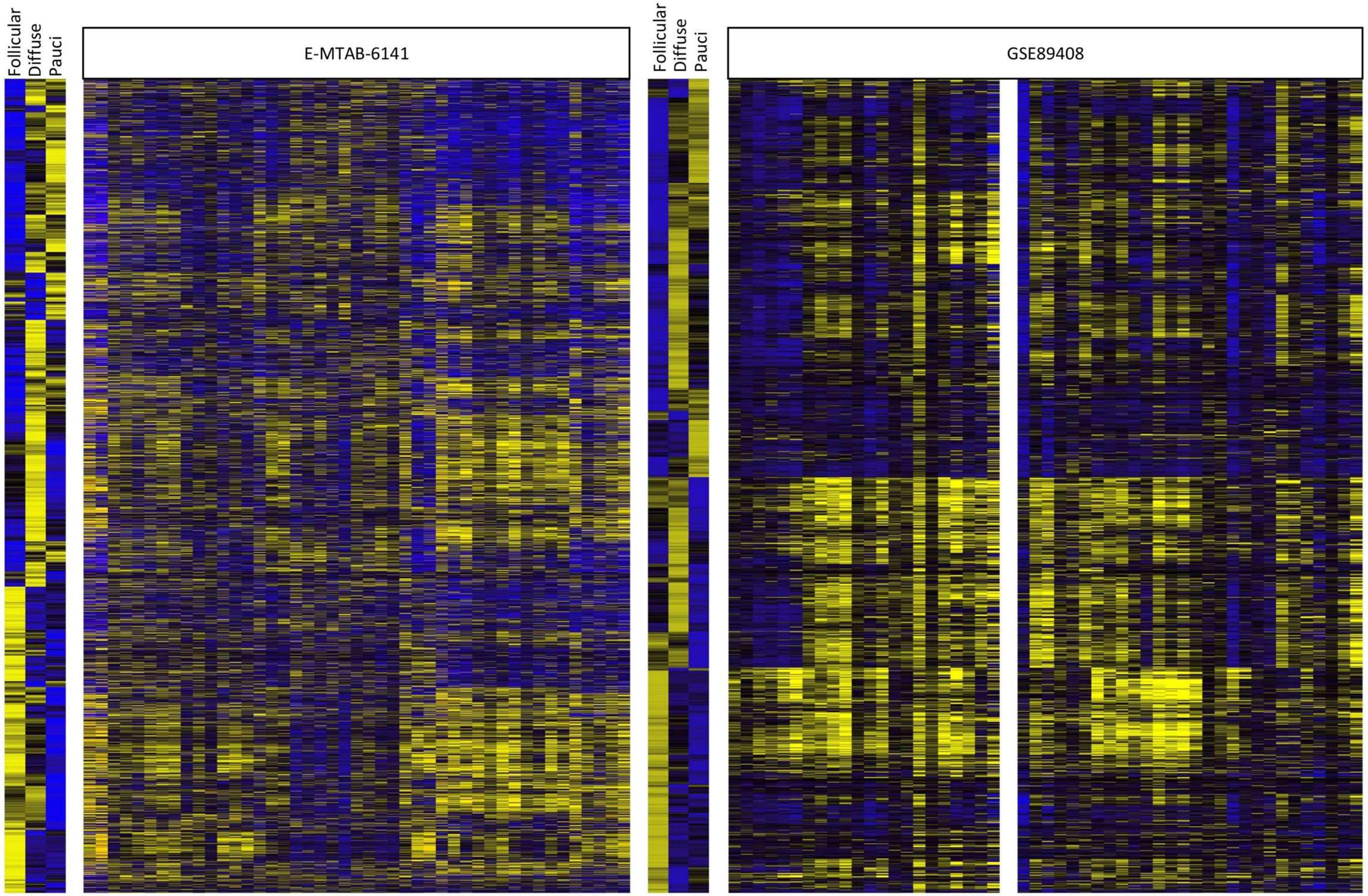


Figure 9.3: Follicular samples from the RNA-seq based datasets E-MTAB-6141 and GSE89408 (early and established).

Due to differences with genes mapped between the microarray and RNA-seq datasets, the archetype derived from GDE48780 is plotted to allow comparisons across all the datasets. Both RNA-seq datasets show strong expression of lymphoid markers identified in the archetype, but like the other datasets in figure 9.2 also demonstrate some overlap with the Diffuse archetype

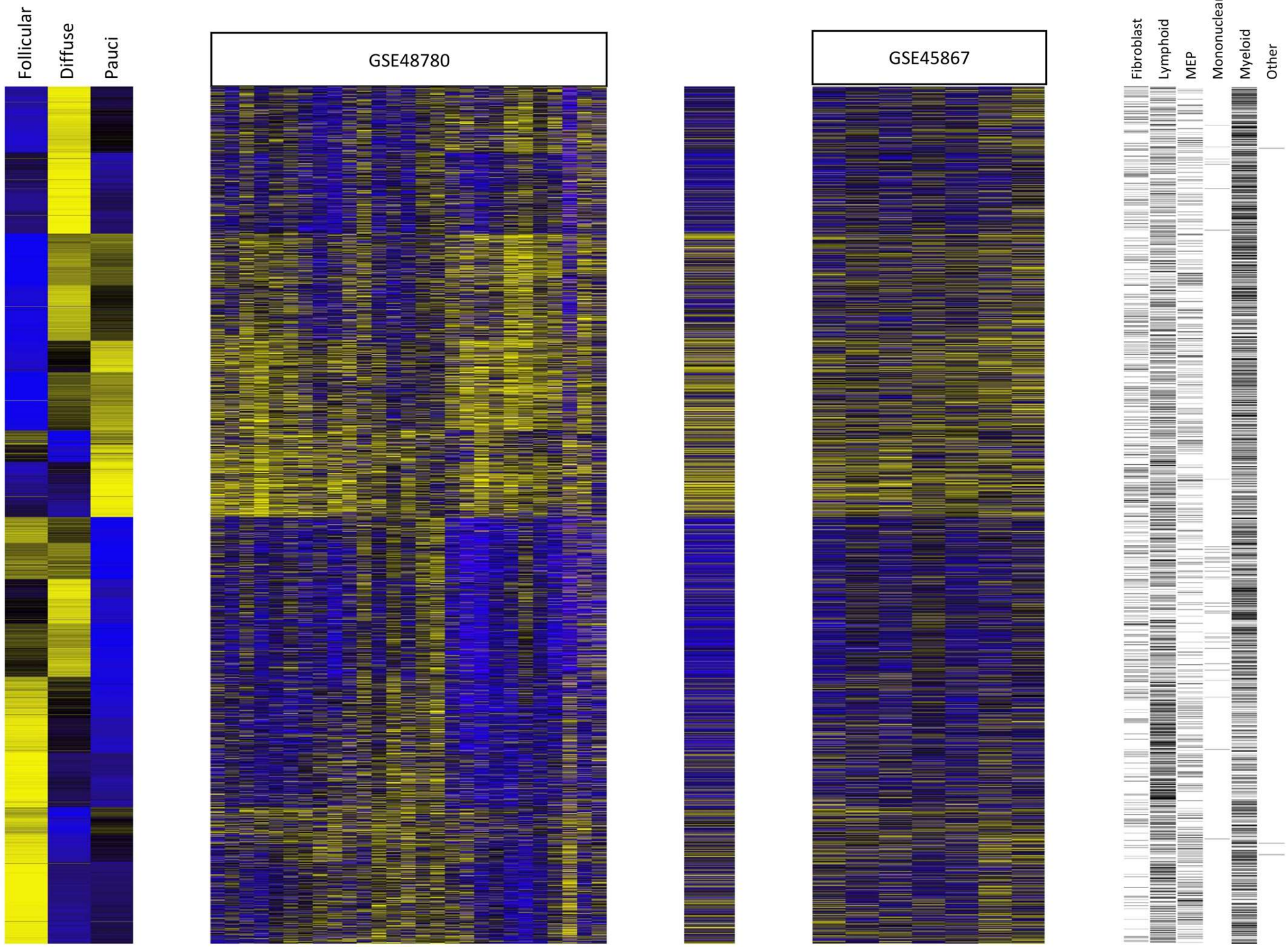


Figure 9.4: Pauci-immune samples from the microarray based datasets GSE48780, GSE24742, and GSE45867.

GSE24742 contains only single sample that is pauci-immune. Pauci-immune samples demonstrate low expression of immune markers in general, and expression is consistent across the datasets

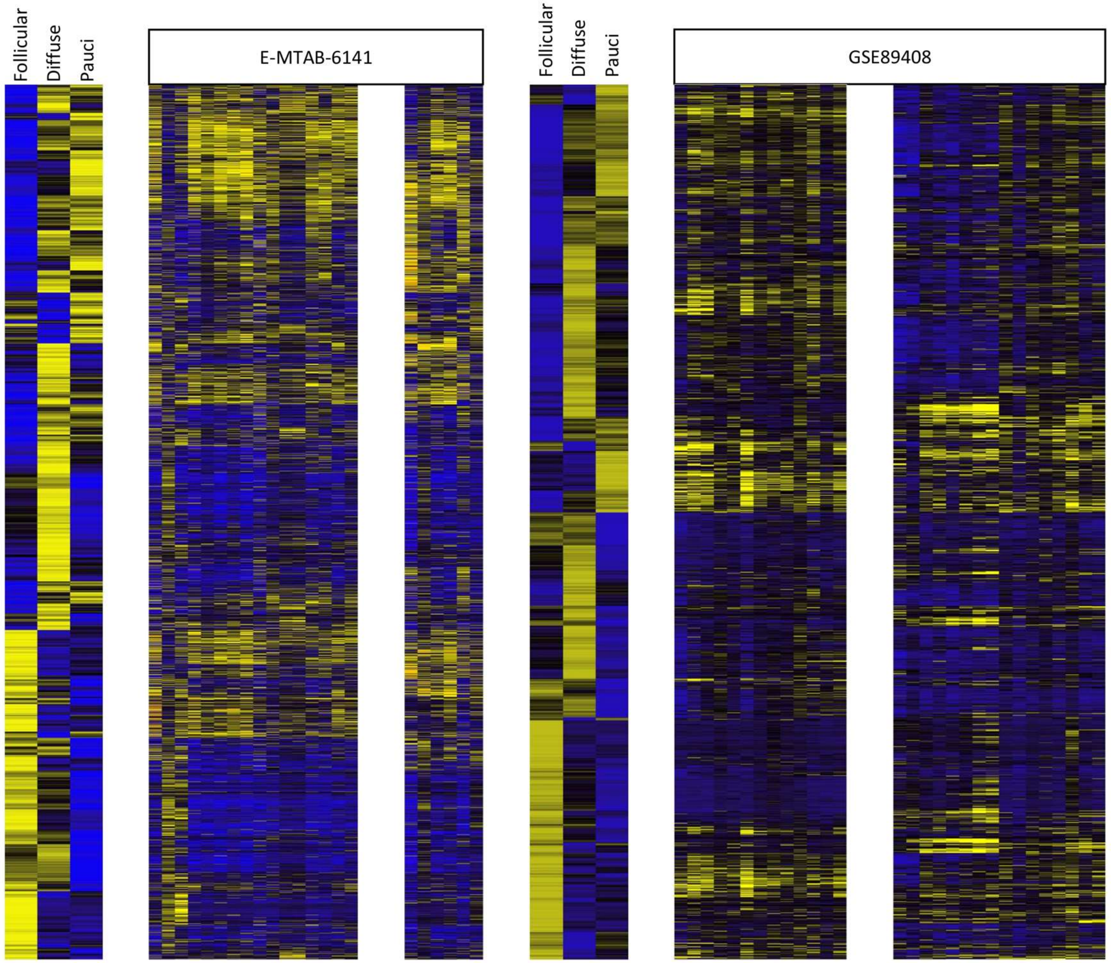


Figure 9.5: Pauci-immune samples from the RNA-seq based datasets E-MTAB-6141 (including ungraded) and GSE89408 (early and established).

Both RNA-seq datasets show good replication of the archetype across the samples, however, early stage disease does show some limited expression of lymphoid and myeloid markers that may reflect the initial inflammation and transient infiltration of immune cells.

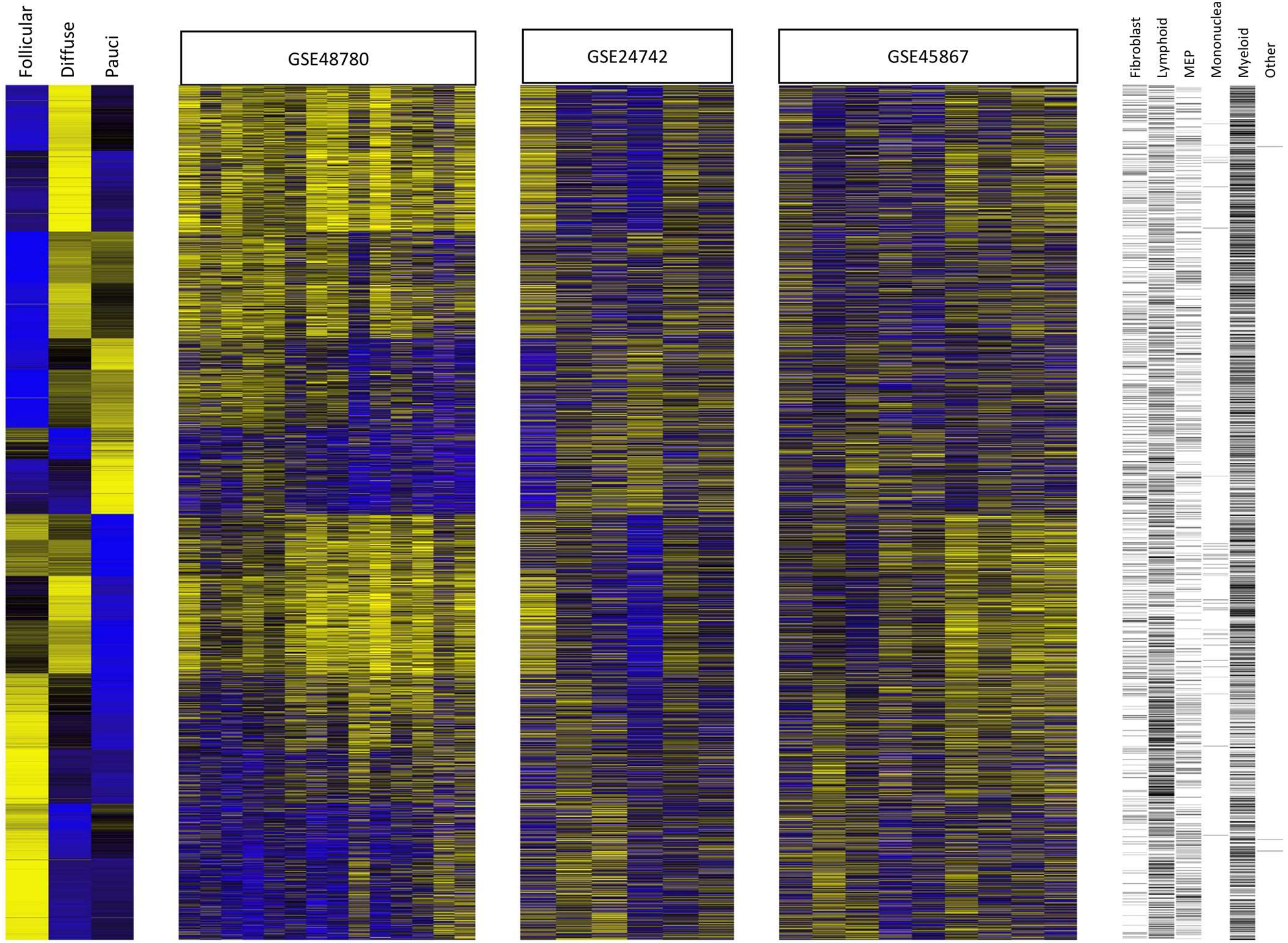


Figure 9.6: Diffuse samples from the microarray based datasets GSE48780, GSE24742, and GSE45867.

Diffuse samples demonstrate considerable variability in transcriptional profile between the datasets.

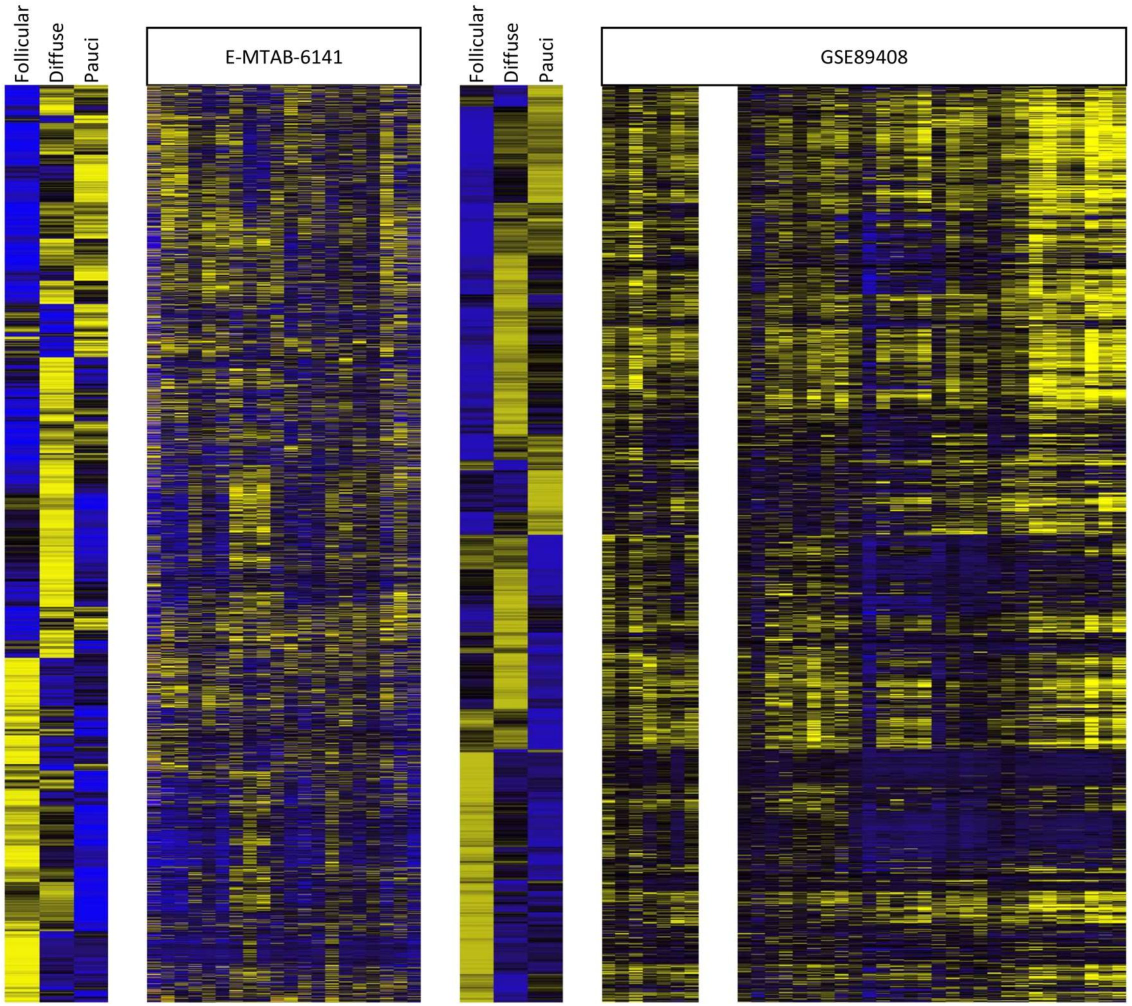


Figure 9.7: Diffuse samples from the RNA-seq based datasets E-MTAB-6141 and GSE89408 (early and established).

The variability in transcriptional profiles is also seen in the RNA-seq datasets, with early disease in GSE89408 shows expression of all markers associated with all 3 pathologies.

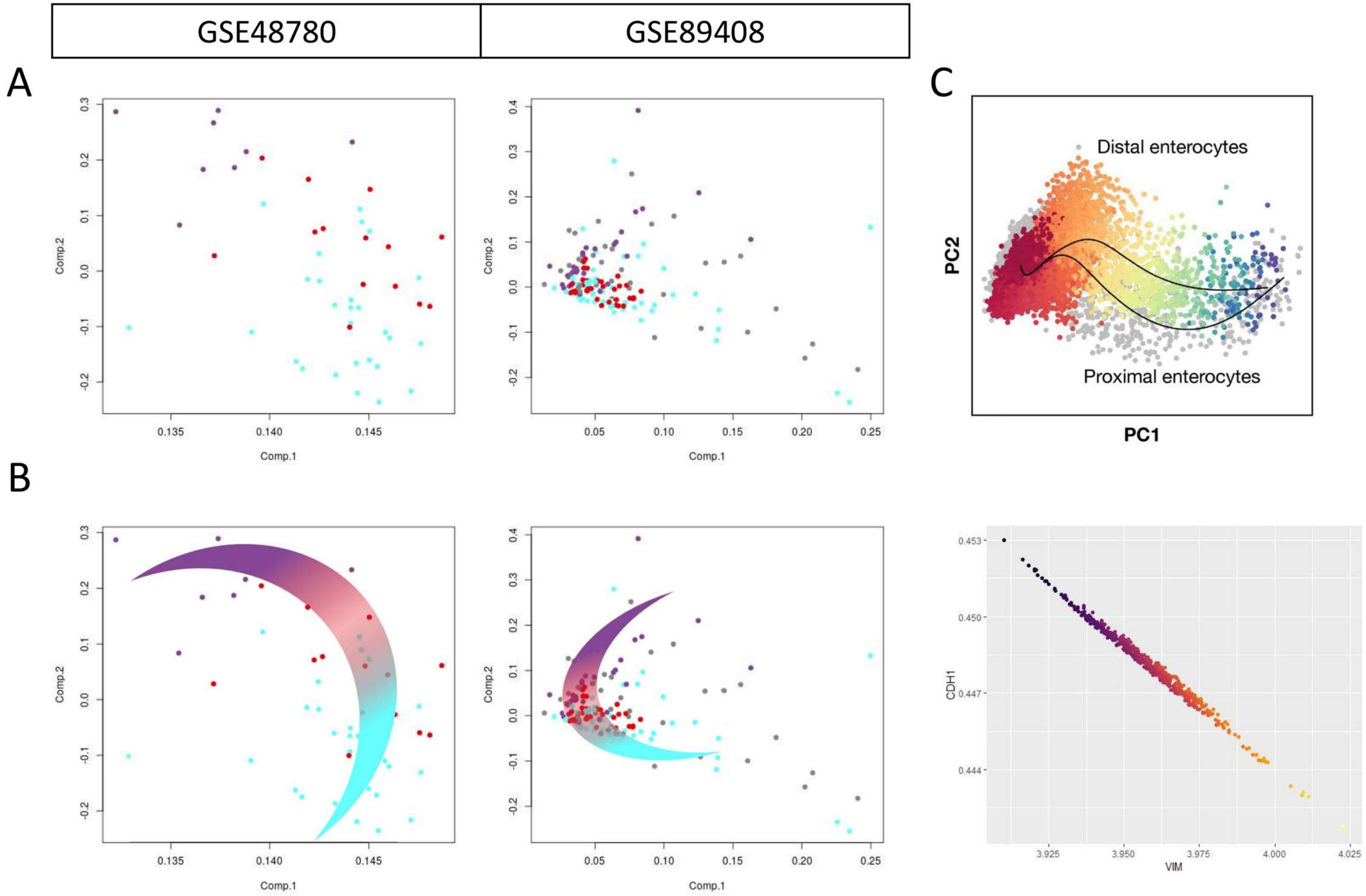


Figure 9.8: Dimensional reduction (PCA) reveals manifold like structures with diffuse samples being midway between follicular and pauci-immune.

***A** shows the raw PCA, with **B** annotating a cartoon manifold above it. **C** shows example data from single cell experiments looking at differentiation pathways: the first looking at mouse intestinal epithelium (Current best practices in single-cell RNA-seq analysis: a tutorial) and the second for epithelial to mesenchymal transition (MAGIC: A diffusion-based imputation method reveals gene-gene interactions in single-cell RNA-sequencing data)*

Thus, diffuse synovitis may reflect an entire spectrum of disease presentations or sub-pathologies.

When looking at how these pathologies behave relative to each other (e.g., using dimensional reduction techniques such as principle component analysis PCA), samples with diffuse synovitis sit between follicular and pauci-immune synovitis. In the field of single-cell genomics, we often see trajectories of cells as they develop towards their fate. It is, therefore, possible that we are seeing a similar behaviour in diffuse synovitis where follicular and pauci-immune synovitis reflect the extremes, and diffuse synovitis the transitionary phases that reside between. Support for this hypothesis is shown in Figure 9.8, which plots the PCA of several of the datasets evaluated in this thesis together with a manifold behaviour of transitioning cells from single-cell sequencing. Whilst this type of analysis requires further exploration, the principles suggest that future studies may require a more detailed interrogation of synovial biopsies. For example, the analysis of sequential biopsies from the same joint or studies in explant cultures to track disease progression.

With the datasets exploring therapeutic responses (GSE24742 & GSE45867) and disease and disease progression (GSE89408), this allows investigation of these transcriptional profiles over time. This reveals an interesting observation that the transcriptional profile of pauci-immune is generally consistent with that of healthy synovium (Figure 9.9), and that following successful treatment those patients with other pathologies take on a similar profile (Figures 7.1, 7.2, 8.3).

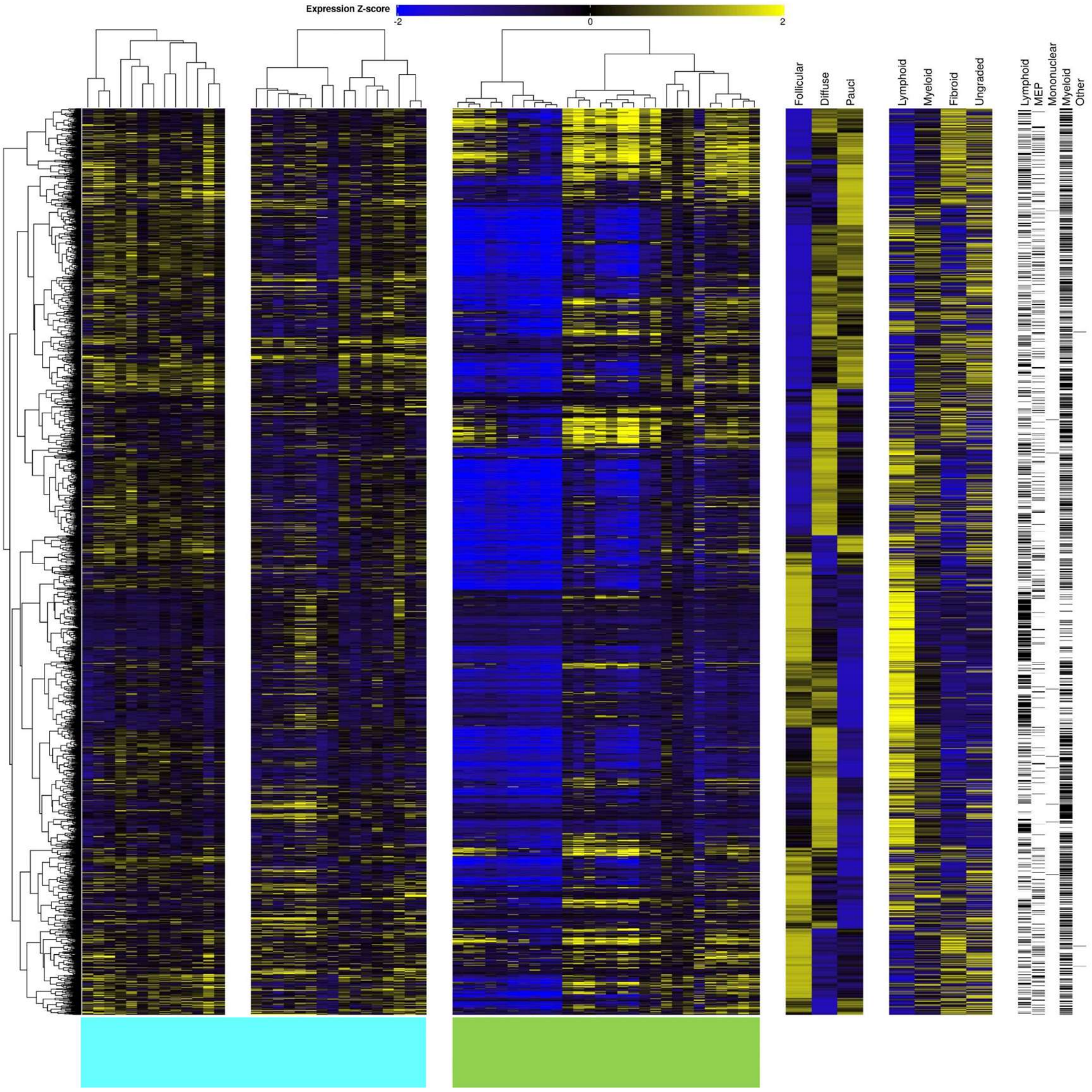


Figure 9.9: Transcriptional profile of Pauci-immune and Healthy synovium.

GSE89408 contains samples from multiple arthritic diseases and healthy synovium, Pauci-immune (Cyan) demonstrates a similar transcriptional profile to healthy synovium (green).

9.3. Missing metadata and underpowered studies.

One issue identified during the course of my studies was the absence of consistent metadata to support the interpretations. The PEAC dataset provides an example of excellent metadata, providing not only histological data, but copious amounts of disease activity measures. This, however, was not always the case. Several of the datasets examined had limited, and often incomplete, metadata. Moreover, the available metadata was often not directly linked to the transcriptomic data. This necessitated the need to extraction additional information from supplementary data within the original publications.

Furthermore, many of these studies have low numbers of samples. GSE24742 has only 12 patients (before and after rituximab), GSE45867 has 20 patients split into 2 groups of therapy 8 methotrexate and 12 tocilizumab. Other datasets had fewer patients, but their focus was usually on comparing rheumatoid arthritis to osteoarthritis or similar analysis. This lack of Power made the generation of prediction tools and classifiers challenging and hampered the analysis of diffuse synovitis due to the complexity of this disease setting. Whilst the impact of batch effects reduced my ability to work with combined datasets. Despite these challenges, I believe that the computational methods generated in this thesis offer real potential for further development as classifiers of synovitis. In this regard, the approaches used provide important proof-of-concept discoveries showcasing how my research may help interrogate the responses to therapy and the pathways driving synovitis. Further advances in machine-based learning will help to refine these approaches. However, my thesis has for the first time described the generation of a bespoke classifier tool for the study of disease heterogeneity in immune-mediated inflammatory disease. I believe that these methods will have real-life utility in future clinal trials.

9.4. Future work.

Data presented in this thesis demonstrate that it is possible to stratify patients based off a small gene signature. However further refinements to these signatures are required to enable them to be introduce into clinical studies and routine clinical practice.

- **Predictive assessment of classifier performance:** the classifiers discussed in this thesis have been assessed using independent clinical datasets. However, ideally these signatures require continued testing (and potentially further honing) using newer cohorts with more complete metadata. Training with newer studies should also allow for refinement of the signature, improving its classification performance. Thus, clinical studies may need to be built around the testing of the tools.
- **Better define the diffuse pathology:** Identifying a consistent transcriptional profile associated with the diffuse pathology is essential for future work. As demonstrated in this thesis, the diffuse samples have a mixed profile of immunological marker expression and this results in significant differences between training and testing datasets. These significantly impacted the classifiers ability to call patients with diffuse pathology. With transcriptional datasets obtained from larger cohorts it should be possible to identify a more robust transcriptional profile that discriminates the diffuse pathology. Possibly linked to smaller sub-groups of diffuse pathology. Moreover, it would be interesting to evaluate the temporal evolution of diffuse pathology to establish whether these forms of synovitis represent defined end-points of disease or specific stages of disease development.

- **Identify an accurate diffuse classifier:** With a better-defined diffuse cohort, retraining the classifiers should allow for improved performance of the gene signatures to stratify these diffuse patients.
- **Therapeutic response signature:** The reporting of larger clinical studies examining the therapeutic responses of patients with distinct forms of synovitis (e.g., R4RA, STRAP studies) has necessitated the need to develop new methodologies that predict disease outcome. Whilst clinical experience offers some helpful insights (e.g. patients with pauci-immune synovitis show poor efficacy to anti-TNFα therapy), this is not a precise science. Therefore, tools (such as those developed in this thesis) offer really opportunities to improve precision medical decision-making. The identification of precision medicine approaches will improve decisions on the best course of therapy sooner in the clinical management and improve real-world opportunities for disease remission in clinical practice. Thus, leading to improvements in patient quality of life, health economics and the overall strain on NHS resources.
- **Identification of the role of STAT1 and STAT3 in synovitis:** As shown in Chapter 8, STAT1 and STAT3 have a clear role in disease progression. However, attempts in this thesis failed to identify the comprehensive list of STAT-associated genes in synovitis. Repeating the ChIP-seq (or adopting ATAC-seq methods) experiments with alternative antibodies may reveal the genes associated with either STAT1 or STAT3 under the conditions of inflammatory arthritis and give insights into the mechanisms that promote disease onset and progression.
- **Define the minimum standard of metadata for studies:** To maximise the utilisation of datasets generated, some thought is required to implement a clinical version of Minimum Information About a Microarray Experiment (MAIME)(225) or Findability, Accessibility, Interoperability, and Reusability guidelines(FAIR)(226), which standardises the clinical metadata format for further analysis and re-evaluation. This would need to be driven by the clinical and academic communities to push for relevant metadata selection and updating records wherever possible

9.5. Conclusion. –

These gene signatures offer some possibilities to improve clinical approaches to treatment in rheumatoid arthritis. Whist evidence for pathology-specific therapeutic responses is limited at the time being(71,72,89,91), ongoing trials are starting to show the association(169). Therefore, with the utilisation of precision medicine this will increase the identification of these associations and improve therapeutic targeting and therefore improve clinical outcomes.

Moreover, whilst this thesis has been focussed on rheumatoid arthritis, the techniques utilised are disease agnostic and therefore can be adapted to other conditions where subclassification of pathology may be needed.

10. Appendix

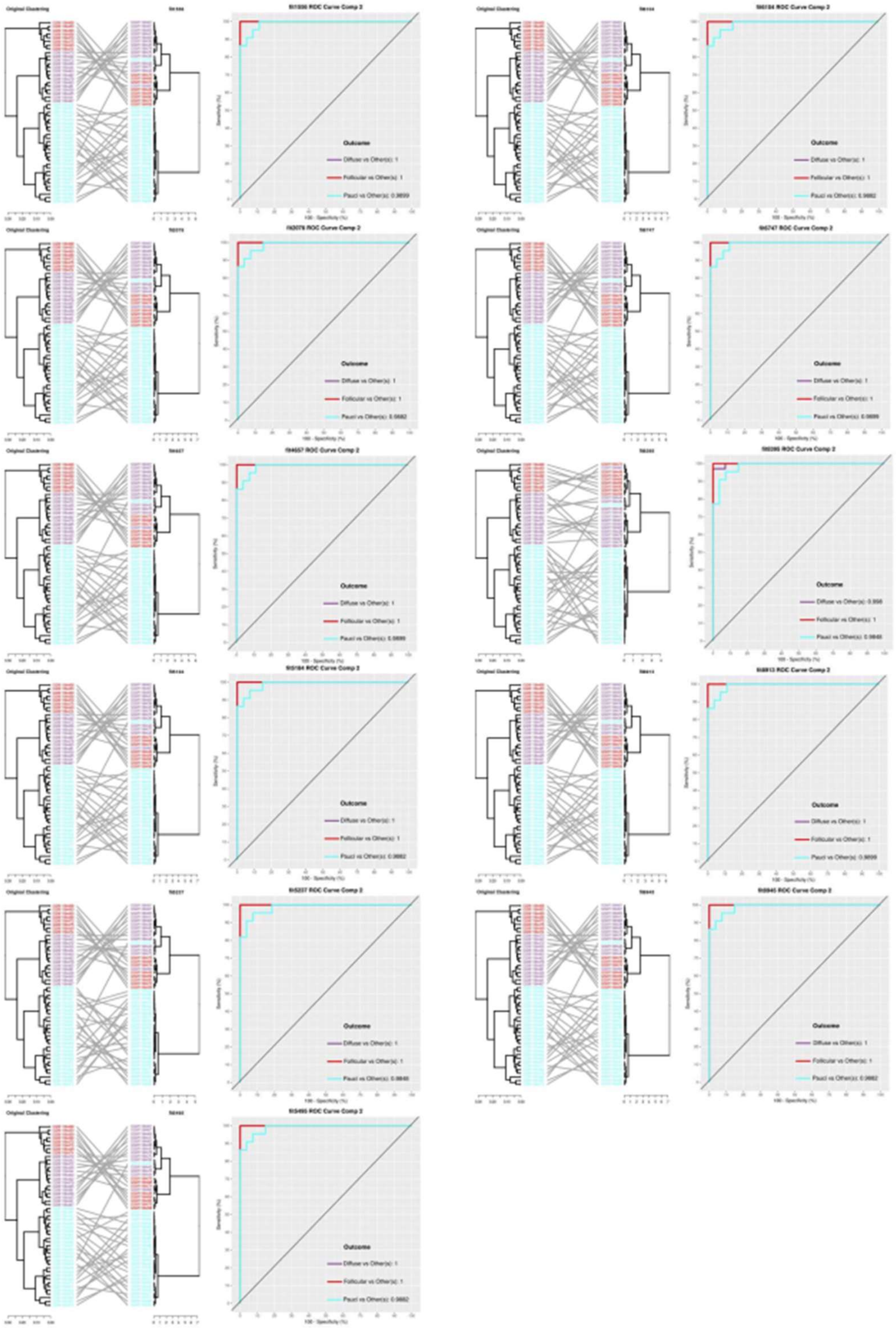


Figure 10.1: Multiple good fit models as determined by ability to replicate the clustering from the original.

Quantification of the performance of the models was first assessed using the tanglegrams, looking for distinct clusters of samples, so minimal inclusion of other pathotypes within the cluster, and that the clusters are distinct for the pathologies (so avoiding clusters of pathotype split across two branches). Performance was also examined by looking at the AUC for the different pathologies when utilising both component 1 and 2

References.

1. The Human Genome Project [Internet]. [cited 2020 Nov 3]. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6875757/

2. Gibbs RA. The Human Genome Project changed everything. Nature Reviews Genetics. 2020 Oct;21(10):575–6.

3. Lander ES. Initial impact of the sequencing of the human genome. Nature. 2011 Feb;470(7333):187–97.

4. Human Whole-Genome Sequencing with the NovaSeq 6000 Sequencing System. :2.

5. Genomics took a long time to fulfil its promise. The Economist [Internet]. 2020 Mar 12 [cited 2020 Nov 3]; Available from: https://www.economist.com/technology-quarterly/2020/03/12/genomics-took-a-long-time-to-fulfil-its-promise

6. Oyelade J, Isewon I, Oladipupo F, Aromolaran O, Uwoghiren E, Ameh F, et al. Clustering Algorithms: Their Application to Gene Expression Data. Bioinform Biol Insights. 2016 Nov 30;10:237–53.

7. Pirim H, Ekşioğlu B, Perkins A, Yüceer Ç. Clustering of High Throughput Gene Expression Data. Comput Oper Res. 2012 Dec;39(12):3046–61.

8. Kaufman L, Rousseeuw PJ. Finding Groups in Data: An Introduction to Cluster Analysis. John Wiley & Sons; 2009. 369 p.

9. Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Research. 2015;43(7):e47.

10. Tibshirani R, Hastie T, Narasimhan B, Chu G. Diagnosis of multiple cancer types by shrunken centroids of gene expression. Proceedings of the National Academy of Sciences. 2002 May 14;99(10):6567–72.

11. Barker M, Rayens W. Partial least squares for discrimination. Journal of Chemometrics. 2003;17(3):166–73.

12. Chung D, Keles S. Sparse Partial Least Squares Classification for High Dimensional Data. Statistical Applications in Genetics and Molecular Biology [Internet]. 2010 Mar 3 [cited 2020 Nov 9];9(1). Available from: https://www.degruyter.com/view/journals/sagmb/9/1/article-sagmb.2010.9.1.1492.xml.xml

13. Ruiz-Perez D, Guan H, Madhivanan P, Mathee K, Narasimhan G. So you think you can PLS-DA? [Internet]. Bioinformatics; 2017 Oct [cited 2020 Oct 18]. Available from: http://biorxiv.org/lookup/doi/10.1101/207225

14. Rohart F, Mason EA, Matigian N, Mosbergen R, Korn O, Chen T, et al. A molecular classification of human mesenchymal stromal cells. PeerJ. 2016;4:e1845.

15. Costa RB, Kurra G, Greenberg L, Geyer CE. Efficacy and cardiac safety of adjuvant trastuzumab-based chemotherapy regimens for HER2-positive early breast cancer. Ann Oncol. 2010 Nov;21(11):2153–60.

16. Choy EH, Kavanaugh AF, Jones SA. The problem of choice: current biologic agents and future prospects in RA. Nature Reviews Rheumatology. 2013 Mar;9(3):154–63.

17. Cojocaru M, Cojocaru IM, Silosi I, Vrabie CD, Tanasescu R. Extra-articular Manifestations in Rheumatoid Arthritis. Maedica (Bucur). 2010 Dec;5(4):286–91.

18. Smolen JS, Steiner G. Therapeutic strategies for rheumatoid arthritis. Nat Rev Drug Discov. 2003 Jun;2(6):473–88.

19. McAllister K, Eyre S, Orozco G. Genetics of rheumatoid arthritis: GWAS and beyond. Open Access Rheumatol. 2011 Jun 7;3:31–46.

20. Smolen JS, Aletaha D, Koeller M, Weisman MH, Emery P. New therapies for treatment of rheumatoid arthritis. Lancet. 2007 Dec 1;370(9602):1861–74.

21. Eyre S, Orozco G, Worthington J. The genetics revolution in rheumatology: large scale genomic arrays and genetic mapping. Nat Rev Rheumatol. 2017 Jul;13(7):421–32.

22. Klareskog L, Padyukov L, Alfredsson L. Smoking as a trigger for inflammatory rheumatic diseases. Curr Opin Rheumatol. 2007 Jan;19(1):49–54.

23. Getts MT, Miller SD. 99th Dahlem Conference on Infection, Inflammation and Chronic Inflammatory Disorders: Triggering of autoimmune diseases by infections. Clinical & Experimental Immunology. 2010;160(1):15–21.

24. Ngo ST, Steyn FJ, McCombe PA. Gender differences in autoimmune disease. Frontiers in Neuroendocrinology. 2014 Aug 1;35(3):347–69.

25. Alpízar-Rodríguez D, Pluchino N, Canny G, Gabay C, Finckh A. The role of female hormonal factors in the development of rheumatoid arthritis. Rheumatology (Oxford). 2017 01;56(8):1254–63.

26. Waaler E. On the Occurrence of a Factor in Human Serum Activating the Specific Agglutination of Sheep Blood Corpuscles. Acta Pathologica Microbiologica Scandinavica. 1940;17(2):172–88.

27. Song YW, Kang EH. The pathogenic role of rheumatoid factor in rheumatoid arthritis. International Journal of Clinical Rheumatology. 2010 Dec;5(6):651–8.

28. de Brito Rocha S, Baldo DC, Andrade LEC. Clinical and pathophysiologic relevance of autoantibodies in rheumatoid arthritis. Advances in Rheumatology. 2019 Jan 17;59(1):2.

29. Derksen VFAM, Huizinga TWJ, van der Woude D. The role of autoantibodies in the pathophysiology of rheumatoid arthritis. Semin Immunopathol. 2017 Jun 1;39(4):437–46.

30. Shi J, Stadt LA van de, Levarht EWN, Huizinga TWJ, Hamann D, Schaardenburg D van, et al. Anti-carbamylated protein (anti-CarP) antibodies precede the onset of rheumatoid arthritis. Annals of the Rheumatic Diseases. 2014 Apr 1;73(4):780–3.

31. Juarez M, Bang H, Hammar F, Reimer U, Dyke B, Sahbudin I, et al. Identification of novel antiacetylated vimentin antibodies in patients with early inflammatory arthritis. Ann Rheum Dis. 2016 Jun;75(6):1099–107.

32. Firestein GS, McInnes IB. Immunopathogenesis of Rheumatoid Arthritis. Immunity. 2017 Feb 21;46(2):183–96.

33. Smolen JS, Aletaha D, Barton A, Burmester GR, Emery P, Firestein GS, et al. Rheumatoid arthritis. Nature Reviews Disease Primers. 2018 Feb 8;4(1):1–23.

34. Müller-Ladner U, Kriegsmann J, Franklin BN, Matsumoto S, Geiler T, Gay RE, et al. Synovial fibroblasts of patients with rheumatoid arthritis attach to and invade normal human cartilage when engrafted into SCID mice. Am J Pathol. 1996 Nov;149(5):1607–15.

35. Croft AP, Campos J, Jansen K, Turner JD, Marshall J, Attar M, et al. Distinct fibroblast subsets drive inflammation and damage in arthritis. Nature. 2019 Jun 1;570(7760):246–51.

36. Reece RJ, Canete JD, Parsons WJ, Emery P, Veale DJ. Distinct vascular patterns of early synovitis in psoriatic, reactive, and rheumatoid arthritis. Arthritis Rheum. 1999 Jul;42(7):1481–4.

37. McInnes IB, Buckley CD, Isaacs JD. Cytokines in rheumatoid arthritis — shaping the immunological landscape. Nat Rev Rheumatol. 2016 Jan;12(1):63–8.

38. McInnes IB, Schett G. Cytokines in the pathogenesis of rheumatoid arthritis. Nature Reviews Immunology. 2007 Jun;7(6):429–42.

39. Nalbant S, Birlik AM. Cytokines in Rheumatoid Arthritis (RA). New Developments in the Pathogenesis of Rheumatoid Arthritis [Internet]. 2017 Feb 22 [cited 2020 Oct 18]; Available from: https://www.intechopen.com/books/new-developments-in-the-pathogenesis-of-rheumatoid-arthritis/cytokines-in-rheumatoid-arthritis-ra-

40. Boyadzhieva V, Stoilov N, Ivanova M, Petrova G, Stoilov R. Real World Experience of Disease Activity in Patients With Rheumatoid Arthritis and Response to Treatment With Varios Biologic DMARDs. Front Pharmacol [Internet]. 2018 [cited 2020 Mar 10];9. Available from: https://www.frontiersin.org/articles/10.3389/fphar.2018.01303/full

41. England BR, Tiong BK, Bergman MJ, Curtis JR, Kazi S, Mikuls TR, et al. 2019 Update of the American College of Rheumatology Recommended Rheumatoid Arthritis Disease Activity Measures. Arthritis Care & Research. 2019;71(12):1540–55.

42. van Gestel AM, Haagsma CJ, van Riel PL. Validation of rheumatoid arthritis improvement criteria that include simplified joint counts. Arthritis Rheum. 1998 Oct;41(10):1845–50.

43. Fransen J, van Riel PLCM. The Disease Activity Score and the EULAR Response Criteria. Rheumatic Disease Clinics of North America. 2009 Nov;35(4):745–57.

44. Singh H, Kumar H, Handa R, Talapatra P, Ray S, Gupta V. Use of Clinical Disease Activity Index Score for Assessment of Disease Activity in Rheumatoid Arthritis Patients: An Indian Experience. Arthritis [Internet]. 2011 [cited 2020 Nov 9];2011. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3254008/

45. Smolen JS, Breedveld FC, Schiff MH, Kalden JR, Emery P, Eberl G, et al. A simplified disease activity index for rheumatoid arthritis for use in clinical practice. Rheumatology (Oxford). 2003 Feb 1;42(2):244–57.

46. Boone NW, Sepriano A, Kuy P-H van der, Janknegt R, Peeters R, Landewé RBM. Routine Assessment of Patient Index Data 3 (RAPID3) alone is insufficient to monitor disease activity in rheumatoid arthritis in clinical practice. RMD Open. 2019 Nov 1;5(2):e001050.

47. Parekh K, Taylor WJ. The Patient Activity Scale-II Is a Generic Indicator of Active Disease in Patients with Rheumatic Disorders. The Journal of Rheumatology. 2010 Sep 1;37(9):1932–4.

48. 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 [Internet]. 2019 Apr 3 [cited 2020 Jul 17];5(1). Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6525606/

49. Smolen JS, Landewé 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 Jun;76(6):960–77.

50. Kiely PDW, Deighton C, Dixey J, Ostor AJK, on behalf of the British Society for Rheumatology Standards, Guidelines and Audit Working Group. Biologic agents for rheumatoid arthritis--negotiating the NICE technology appraisals. Rheumatology. 2012 Jan 1;51(1):24–31.

51. Moreland LW, O'Dell JR, Paulus HE, Curtis JR, Bathon JM, Clair EWSt, et al. A Randomized Comparative Effectiveness Study of Oral Triple Therapy versus Etanercept plus Methotrexate in Early, Aggressive Rheumatoid Arthritis. Arthritis Rheum. 2012 Sep;64(9):2824–35.

52. Singh JA, Christensen R, Wells GA, Suarez-Almazor ME, Buchbinder R, Lopez-Olivo MA, et al. Biologics for rheumatoid arthritis: an overview of Cochrane reviews. Cochrane Database Syst Rev. 2009 Oct 7;(4):CD007848.

53. Triaille C, Lauwerys BR. Synovial Tissue: Turning the Page to Precision Medicine in Arthritis. Front Med [Internet]. 2019 [cited 2020 Mar 10];6. Available from: https://www.frontiersin.org/articles/10.3389/fmed.2019.00046/full

54. Pitzalis C, Choy EHS, Buch MH. Transforming clinical trials in rheumatology: towards patient-centric precision medicine. Nature Reviews Rheumatology. 2020 Oct;16(10):590–9.

55. Dey M, Zhao SS, Moots RJ. Anti-TNF biosimilars in rheumatology: the end of an era? Expert Opinion on Biological Therapy. 2020 Jul 31;0(0):1–7.

56. MA X, XU S. TNF inhibitor therapy for rheumatoid arthritis. Biomed Rep. 2013;1(2):177–84.

57. Scott LJ. Tocilizumab: A Review in Rheumatoid Arthritis. Drugs. 2017 Nov;77(17):1865–79.

58. Lacroix M, Rousseau F, Guilhot F, Malinge P, Magistrelli G, Herren S, et al. Novel Insights into Interleukin 6 (IL-6) Cis- and Trans-signaling Pathways by Differentially Manipulating the Assembly of the IL-6 Signaling Complex. J Biol Chem. 2015 Nov 6;290(45):26943–53.

59. Mertens M, Singh JA. Anakinra for rheumatoid arthritis: a systematic review. J Rheumatol. 2009 Jun;36(6):1118–25.

60. Kucharz EJ, Stajszczyk M, Kotulska-Kucharz A, Batko B, Brzosko M, Jeka S, et al. Tofacitinib in the treatment of patients with rheumatoid arthritis: position statement of experts of the Polish Society for Rheumatology. Reumatologia. 2018;56(4):203–11.

61. Vital EM, Emery P. Abatacept in the treatment of rheumatoid arthritis. Ther Clin Risk Manag. 2006 Dec;2(4):365–75.

62. Cohen MD, Keystone E. Rituximab for Rheumatoid Arthritis. Rheumatol Ther. 2015 Aug 19;2(2):99–111.

63. Aletaha D, Smolen JS. Diagnosis and Management of Rheumatoid Arthritis: A Review. JAMA. 2018 Oct 2;320(13):1360–72.

64. Köhler BM, Günther J, Kaudewitz D, Lorenz H-M. Current Therapeutic Options in the Treatment of Rheumatoid Arthritis. J Clin Med [Internet]. 2019 Jun 28 [cited 2020 Nov 9];8(7). Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6678427/

65. Pitzalis C, Kelly S, Humby F. New learnings on the pathophysiology of RA from synovial biopsies. Curr Opin Rheumatol. 2013 May;25(3):334–44.

66. Lauwerys BR, Hernández-Lobato D, Gramme P, Ducreux J, Dessy A, Focant I, et al. Heterogeneity of Synovial Molecular Patterns in Patients with Arthritis. PLoS One [Internet]. 2015 Apr 30 [cited 2019 Oct 28];10(4). Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4415786/

67. Orr C, Vieira-Sousa E, Boyle DL, Buch MH, Buckley CD, Cañete JD, et al. Synovial tissue research: a state-of-the-art review. Nature Reviews Rheumatology. 2017 Aug;13(8):463–75.

68. Wechalekar MD, Smith MD. Arthroscopic guided synovial biopsy in rheumatology: current perspectives. International Journal of Rheumatic Diseases. 2017;20(2):141–4.

69. Wechalekar MD, Smith MD. Utility of arthroscopic guided synovial biopsy in understanding synovial tissue pathology in health and disease states. World J Orthop. 2014 Nov 18;5(5):566–73.

70. Ouboussad L, Burska AN, Melville A, Buch MH. Synovial Tissue Heterogeneity in Rheumatoid Arthritis and Changes With Biologic and Targeted Synthetic Therapies to Inform Stratified Therapy. Front Med [Internet]. 2019 [cited 2020 Mar 10];6. Available from: https://www.frontiersin.org/articles/10.3389/fmed.2019.00045/full

71. Humby F, Lewis M, Ramamoorthi N, Hackney JA, Barnes MR, Bombardieri M, et al. Synovial cellular and molecular signatures stratify clinical response to csDMARD therapy and predict radiographic progression in early rheumatoid arthritis patients. Annals of the Rheumatic Diseases. 2019 Jun 1;78(6):761–72.

72. Nerviani A, Di Cicco M, Mahto A, Lliso-Ribera G, Rivellese F, Thorborn G, et al. A Pauci-Immune Synovial Pathotype Predicts Inadequate Response to TNFα-Blockade in Rheumatoid Arthritis Patients. Front Immunol. 2020 May 5;11:845.

73. Dennis G, Holweg CT, Kummerfeld SK, Choy DF, Setiadi AF, Hackney JA, et al. Synovial phenotypes in rheumatoid arthritis correlate with response to biologic therapeutics. Arthritis Research & Therapy. 2014 Apr 30;16(2):R90.

74. Burska AN, Roget K, Blits M, Soto Gomez L, van de Loo F, Hazelwood LD, et al. Gene expression analysis in RA: towards personalized medicine. The Pharmacogenomics Journal. 2014 Apr;14(2):93–106.

75. Aletaha D, Neogi T, Silman AJ, Funovits J, Felson DT, Bingham CO, et al. 2010 Rheumatoid arthritis classification criteria: An American College of Rheumatology/European League Against Rheumatism collaborative initiative. Arthritis & Rheumatism. 2010 Sep;62(9):2569–81.

76. van der Helm-van Mil AHM, Huizinga TWJ. The 2010 ACR/EULAR criteria for rheumatoid arthritis: do they affect the classification or diagnosis of rheumatoid arthritis? Ann Rheum Dis. 2012 Oct;71(10):1596–8.

77. Cuppen BVJ, Welsing PMJ, Sprengers JJ, Bijlsma JWJ, Marijnissen ACA, van Laar JM, et al. Personalized biological treatment for rheumatoid arthritis: a systematic review with a focus on clinical applicability. Rheumatology (Oxford). 2016 May;55(5):826–39.

78. Devauchelle V, Marion S, Cagnard N, Mistou S, Falgarone G, Breban M, et al. DNA microarray allows molecular profiling of rheumatoid arthritis and identification of pathophysiological targets. Genes & Immunity. 2004 Dec;5(8):597–608.

79. Pratt AG, Swan DC, Richardson S, Wilson G, Hilkens CMU, 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. Ann Rheum Dis. 2012 Aug;71(8):1374–81.

80. Klaasen R, Thurlings RM, Wijbrandts CA, Kuijk AW van, Baeten D, Gerlag DM, et al. The relationship between synovial lymphocyte aggregates and the clinical response to infliximab in rheumatoid arthritis: A prospective study. Arthritis & Rheumatism. 2009;60(11):3217–24.

81. Lindberg J, Wijbrandts CA, van Baarsen LG, Nader G, Klareskog L, Catrina A, et al. The Gene Expression Profile in the Synovium as a Predictor of the Clinical Response to Infliximab Treatment in Rheumatoid Arthritis. Vij N, editor. PLoS ONE. 2010 Jun 25;5(6):e11310.

82. Lindberg J, af Klint E, Catrina A, Nilsson P, Klareskog L, Ulfgren A-K, et al. Effect of infliximab on mRNA expression profiles in synovial tissue of rheumatoid arthritis patients. Arthritis Res Ther. 2006;8(6):R179.

83. Gazeau P, Alegria GC, Devauchelle-Pensec V, Jamin C, Lemerle J, Bendaoud B, et al. Memory B Cells and Response to Abatacept in Rheumatoid Arthritis. Clinic Rev Allerg Immunol. 2017 Oct 1;53(2):166–76.

84. Scarsi M, Ziglioli T, Airo' P. Baseline Numbers of Circulating CD28-negative T Cells May Predict Clinical Response to Abatacept in Patients with Rheumatoid Arthritis. J Rheumatol. 2011 Oct;38(10):2105–11.

85. Bresnihan B, Pontifex E, Thurlings RM, Vinkenoog M, El-Gabalawy H, Fearon U, et al. Synovial Tissue Sublining CD68 Expression Is a Biomarker of Therapeutic Response in Rheumatoid Arthritis Clinical Trials: Consistency Across Centers. J Rheumatol. 2009 Aug;36(8):1800–2.

86. Badot V, Galant C, Nzeusseu Toukap A, Theate I, Maudoux A-L, Van den Eynde BJ, et al. Gene expression profiling in the synovium identifies a predictive signature of absence of

response to adalimumab therapy in rheumatoid arthritis. Arthritis Research & Therapy. 2009;11(2):R57.

87. De Groof A, Ducreux J, Humby F, Nzeusseu Toukap A, Badot V, Pitzalis C, et al. Higher expression of TNFα-induced genes in the synovium of patients with early rheumatoid arthritis correlates with disease activity, and predicts absence of response to first line therapy. Arthritis Research & Therapy. 2016 Jan 20;18(1):19.

88. Jones GW, Jones SA. Ectopic lymphoid follicles: inducible centres for generating antigen-specific immune responses within tissues. Immunology. 2016 Feb;147(2):141–51.

89. Nakamura S, Suzuki K, Iijima H, Hata Y, Lim CR, Ishizawa Y, et al. Identification of baseline gene expression signatures predicting therapeutic responses to three biologic agents in rheumatoid arthritis: a retrospective observational study. Arthritis Research & Therapy. 2016 Jul 19;18(1):159.

90. Vieira-Sousa E, Gerlag DM, Tak PP. Synovial Tissue Response to Treatment in Rheumatoid Arthritis. The Open Rheumatology Journal [Internet]. 2011 Dec 30 [cited 2020 Oct 29];5(1). Available from: https://openrheumatologyjournal.com/VOLUME/5/PAGE/115/FULLTEXT/

91. Tony H-P, Roll P, Mei HE, Blümner E, Straka A, Gnuegge L, et al. Combination of B cell biomarkers as independent predictors of response in patients with rheumatoid arthritis treated with rituximab. Clin Exp Rheumatol. 2015 Dec;33(6):887–94.

92. Smith SL, Plant D, Eyre S, Hyrich K, Morgan AW, Wilson AG, et al. The predictive value of serum S100A9 and response to etanercept is not confirmed in a large UK rheumatoid arthritis cohort. Rheumatology (Oxford). 2017 Jun 1;56(6):1019–24.

93. Lliso-Ribera G, Humby F, Lewis M, Nerviani A, Mauro D, Rivellese F, et al. Synovial tissue signatures enhance clinical classification and prognostic/treatment response algorithms in early inflammatory arthritis and predict requirement for subsequent biological therapy: results from the pathobiology of early arthritis cohort (PEAC). Annals of the Rheumatic Diseases. 2019 Dec 1;78(12):1642–52.

94. Mulhearn B, Barton A, Viatte S. Using the Immunophenotype to Predict Response to Biologic Drugs in Rheumatoid Arthritis. J Pers Med. 2019 Oct 2;9(4).

95. Pratt AG, Brown PM, Cockell SJ, Wilson G, Isaacs JD. A3.2 A CD4+ T-Cell Gene Expression Signature Predicts Drug Survival on Methotrexate Monotherapy in Early Rheumatoid Arthritis. Annals of the Rheumatic Diseases. 2013 Mar 1;72(Suppl 1):A13–4.

96. Mans K, Tandon N, Sohnrey C, Bolle S, Grützkau A, Burmester GR, et al. A7.17 Microarray Gene Expression Profiling of Rheumatoid Arthritis Patients for Prediction of Response to Methotrexate Treatment. Annals of the Rheumatic Diseases. 2013 Mar 1;72(Suppl 1):A54–A54.

97. van der Pouw Kraan TC, Wijbrandts CA, van Baarsen LG, Rustenburg F, Baggen JM, Verweij CL, et al. Responsiveness to anti-tumour necrosis factor alpha therapy is related to pre-treatment tissue inflammation levels in rheumatoid arthritis patients. Annals of the Rheumatic Diseases. 2008 Apr;67(4):563–6.

98. Sekiguchi N, Kawauchi S, Furuya T, Inaba N, Matsuda K, Ando S, et al. Messenger ribonucleic acid expression profile in peripheral blood cells from RA patients following

treatment with an anti-TNF-α monoclonal antibody, infliximab. Rheumatology (Oxford). 2008 Jun 1;47(6):780–8.

99. Tanino M, Matoba R, Nakamura S, Kameda H, Amano K, Okayama T, et al. Prediction of efficacy of anti-TNF biologic agent, infliximab, for rheumatoid arthritis patients using a comprehensive transcriptome analysis of white blood cells. Biochem Biophys Res Commun. 2009 Sep 18;387(2):261–5.

100. Szekanecz Z, Meskó B, Poliska S, Váncsa A, Palatka K, Holló Z, et al. A7.20 Response to Infliximab Therapy can be Predicted using Distinct, Non-Overlapping Gene Panels of Peripheral Blood Gene Expression in Rheumatoid Arthritis and Crohn's Disease. Annals of the Rheumatic Diseases. 2013 Mar 1;72(Suppl 1):A55–A55.

101. Julià A, Erra A, Palacio C, Tomas C, Sans X, Barceló P, et al. An Eight-Gene Blood Expression Profile Predicts the Response to Infliximab in Rheumatoid Arthritis. PLoS One [Internet]. 2009 Oct 22 [cited 2020 Nov 10];4(10). Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2762038/

102. Julià A, Barceló M, Erra A, Palacio C, Marsal S. Identification of candidate genes for rituximab response in rheumatoid arthritis patients by microarray expression profiling in blood cells. Pharmacogenomics. 2009 Oct;10(10):1697–708.

103. Raterman HG, Vosslamber S, de Ridder S, Nurmohamed MT, Lems WF, Boers M, et al. The interferon type I signature towards prediction of non-response to rituximab in rheumatoid arthritis patients. Arthritis Res Ther. 2012;14(2):R95.

104. Mesko B, Poliska S, Szamosi S, Szekanecz Z, Podani J, Varadi C, et al. Peripheral blood gene expression and IgG glycosylation profiles as markers of tocilizumab treatment in rheumatoid arthritis. J Rheumatol. 2012 May;39(5):916–28.

105. Hammer HB, Fagerhol MK, Wien TN, Kvien TK. The soluble biomarker calprotectin (an S100 protein) is associated to ultrasonographic synovitis scores and is sensitive to change in patients with rheumatoid arthritis treated with adalimumab. Arthritis Res Ther. 2011;13(5):R178.

106. Foell D, Kane D, Bresnihan B, Vogl T, Nacken W, Sorg C, et al. Expression of the pro-inflammatory protein S100A12 (EN-RAGE) in rheumatoid and psoriatic arthritis. Rheumatology (Oxford). 2003 Nov;42(11):1383–9.

107. Jones GW, McLoughlin RM, Hammond VJ, Parker CR, Williams JD, Malhotra R, et al. Loss of CD4 $^{+}$ T Cell IL-6R Expression during Inflammation Underlines a Role for IL-6 *Trans* Signaling in the Local Maintenance of Th17 Cells. The Journal of Immunology. 2010 Feb 15;184(4):2130–9.

108. Yoshida H, Hamano S, Senaldi G, Covey T, Faggioni R, Mu S, et al. WSX-1 Is Required for the Initiation of Th1 Responses and Resistance to L. major Infection. Immunity. 2001 Oct 1;15(4):569–78.

109. Jones GW, Hill DG, Sime K, Williams AS. In Vivo Models for Inflammatory Arthritis. In: Jenkins BJ, editor. Inflammation and Cancer: Methods and Protocols [Internet]. New York, NY: Springer New York; 2018 [cited 2019 Jun 21]. p. 101–18. (Methods in Molecular Biology). Available from: https://doi.org/10.1007/978-1-4939-7568-6_9

110. Kollias G, Papadaki P, Apparailly F, Vervoordeldonk MJ, Holmdahl R, Baumans V, et al. Animal models for arthritis: innovative tools for prevention and treatment. Annals of the Rheumatic Diseases. 2011 Aug 1;70(8):1357–62.

111. Brackertz D, Mitchell GF, Mackay IR. Antigen-induced arthritis in mice. Arthritis & Rheumatism. 1977;20(3):841–50.

112. Brackertz D, Mitchell GF, Vadas MA, Mackay IR, Miller JFAP. Studies on Antigen-Induced Arthritis in Mice: II. Immunologic Correlates of Arthritis Susceptibility in Mice. The Journal of Immunology. 1977 May 1;118(5):1639–44.

113. Solomon MJ, Varshavsky A. Formaldehyde-mediated DNA-protein crosslinking: a probe for in vivo chromatin structures. Proc Natl Acad Sci U S A. 1985 Oct;82(19):6470–4.

114. Gilmour DS, Lis JT. In vivo interactions of RNA polymerase II with genes of Drosophila melanogaster. Mol Cell Biol. 1985 Aug;5(8):2009–18.

115. Gilmour DS, Lis JT. Detecting protein-DNA interactions in vivo: distribution of RNA polymerase on specific bacterial genes. Proc Natl Acad Sci U S A. 1984 Jul;81(14):4275–9.

116. Euskirchen GM, Rozowsky JS, Wei C-L, Lee WH, Zhang ZD, Hartman S, et al. Mapping of transcription factor binding regions in mammalian cells by ChIP: Comparison of array- and sequencing-based technologies. Genome Res. 2007 Jun;17(6):898–909.

117. Zhou Y, Kurukuti S, Saffrey P, Vukovic M, Michie AM, Strogantsev R, et al. Chromatin looping defines expression of TAL1, its flanking genes, and regulation in T-ALL. Blood. 2013 Dec 19;122(26):4199–209.

118. Gautier L, Cope L, Bolstad BM, Irizarry RA. affy—analysis of Affymetrix GeneChip data at the probe level. Bioinformatics. 2004;20(3):307–315.

119. McCall MN, Bolstad BM, Irizarry RA. Frozen robust multiarray analysis (fRMA). Biostatistics. 2010 Apr 1;11(2):242–53.

120. Zhang Y, Jenkins DF, Manimaran S, Johnson WE. Alternative empirical Bayes models for adjusting for batch effects in genomic studies. BMC Bioinformatics. 2018 Jul 13;19(1):262.

121. Gaujoux R, Seoighe C. CellMix: a comprehensive toolbox for gene expression deconvolution. Bioinformatics. 2013 Sep 1;29(17):2211–2.

122. Abbas AR, Wolslegel K, Seshasayee D, Modrusan Z, Clark HF. Deconvolution of Blood Microarray Data Identifies Cellular Activation Patterns in Systemic Lupus Erythematosus. PLoS One [Internet]. 2009 Jul 1 [cited 2017 Sep 15];4(7). Available from: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2699551/

123. Abbas AR, Baldwin D, Ma Y, Ouyang W, Gurney A, Martin F, et al. Immune response in silico (IRIS): immune-specific genes identified from a compendium of microarray expression data. Genes Immun. 2005 Mar 24;6(4):319–31.

124. Grigoryev YA, Kurian SM, Avnur Z, Borie D, Deng J, Campbell D, et al. Deconvoluting post-transplant immunity: cell subset-specific mapping reveals pathways for activation and expansion of memory T, monocytes and B cells. PLoS ONE. 2010 Oct 14;5(10):e13358.

125. Watkins NA, Gusnanto A, de Bono B, De S, Miranda-Saavedra D, Hardie DL, et al. A HaemAtlas: characterizing gene expression in differentiated human blood cells. Blood. 2009 May 7;113(19):e1–9.

126. Filer A, Antczak P, Parsonage GN, Legault HM, O'Toole M, Pearson MJ, et al. Stromal Transcriptional Profiles Reveal Hierarchies of Anatomical Site, Serum Response and Disease and Identify Disease Specific Pathways. PLOS ONE. 2015 Mar 25;10(3):e0120917.

127. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics. 2009 Jul 15;25(14):1754–60.

128. Zhang Y, Liu T, Meyer CA, Eeckhoute J, Johnson DS, Bernstein BE, et al. Model-based analysis of ChIP-Seq (MACS). Genome Biol. 2008;9(9):R137.

129. Huang W, Loganantharaj R, Schroeder B, Fargo D, Li L. PAVIS: a tool for Peak Annotation and Visualization. Bioinformatics. 2013 Dec 1;29(23):3097–9.

130. Lewis MJ, Barnes MR, Blighe K, Goldmann K, Rana S, Hackney JA, et al. Molecular Portraits of Early Rheumatoid Arthritis Identify Clinical and Treatment Response Phenotypes. Cell Rep. 2019 Aug 27;28(9):2455–70.

131. Bugatti S, Manzo A, Montecucco C, Caporali R. The Clinical Value of Autoantibodies in Rheumatoid Arthritis. Front Med [Internet]. 2018 [cited 2020 Mar 10];5. Available from: https://www.frontiersin.org/articles/10.3389/fmed.2018.00339/full

132. Smits M, van de Groes S, Thurlings RM. Synovial Tissue Biopsy Collection by Rheumatologists: Ready for Clinical Implementation? Front Med [Internet]. 2019 [cited 2020 Mar 10];6. Available from: https://www.frontiersin.org/articles/10.3389/fmed.2019.00138/full

133. Humby FC. Synovial Tissue Sampling in Rheumatological Practice—Past Developments and Future Perspectives. Front Med [Internet]. 2019 [cited 2020 Mar 10];6. Available from: https://www.frontiersin.org/articles/10.3389/fmed.2019.00004/full

134. Bugatti S, Manzo A, Bombardieri M, Vitolo B, Humby F, Kelly S, et al. Synovial Tissue Heterogeneity and Peripheral Blood Biomarkers. Curr Rheumatol Rep. 2011 Aug 17;13(5):440.

135. Townsend MJ. Molecular and cellular heterogeneity in the Rheumatoid Arthritis synovium: Clinical correlates of synovitis. Best Practice & Research Clinical Rheumatology. 2014 Aug 1;28(4):539–49.

136. van der Pouw Kraan TCTM, van Gaalen FA, Huizinga TWJ, Pieterman E, Breedveld FC, Verweij CL. Discovery of distinctive gene expression profiles in rheumatoid synovium using cDNA microarray technology: evidence for the existence of multiple pathways of tissue destruction and repair. Genes & Immunity. 2003 Apr;4(3):187–96.

137. Baarsen LGM van, Wijbrandts CA, Timmer TCG, Kraan TCTM van der P, Tak PP, Verweij CL. Synovial tissue heterogeneity in rheumatoid arthritis in relation to disease activity and biomarkers in peripheral blood. Arthritis & Rheumatism. 2010;62(6):1602–7.

138. Balanescu A, Wiland P. Maximizing early treatment with biologics in patients with rheumatoid arthritis: the ultimate breakthrough in joints preservation. Rheumatol Int. 2013 Jan 9;33(6):1379–86.

139. Lindstrom TM, Robinson WH. Biomarkers for rheumatoid arthritis: Making it personal. Scand J Clin Lab Invest Suppl. 2010 Jul;242:79–84.

140. Veale DJ. Synovial Tissue Biopsy Research. Front Med [Internet]. 2019 [cited 2020 Mar 10];6. Available from: https://www.frontiersin.org/articles/10.3389/fmed.2019.00072/full

141. van Baarsen LG, Bos CL, Pouw Kraan TC van der, Verweij CL. Transcription profiling of rheumatic diseases. Arthritis Res Ther. 2009;11(1):207.

142. Lindberg J, Ulfgren K, Lundeberg J. Effect of infliximab on mRNA expression profiles in synovial tissue of rheumatoid arthritis patients. Arthritis Research. 8(6):12.

143. Lindberg J, Wijbrandts CA, van Baarsen LG, Nader G, Klareskog L, Catrina A, et al. The Gene Expression Profile in the Synovium as a Predictor of the Clinical Response to Infliximab Treatment in Rheumatoid Arthritis. Vij N, editor. PLoS ONE. 2010 Jun 25;5(6):e11310.

144. Ducreux J, Durez P, Galant C, Toukap AN, Eynde BV den, Houssiau FA, et al. Global Molecular Effects of Tocilizumab Therapy in Rheumatoid Arthritis Synovium. Arthritis & Rheumatology. 2014;66(1):15–23.

145. Gutierrez-Roelens I, Galant C, Theate I, Lories RJ, Durez P, Nzeusseu-Toukap A, et al. Rituximab treatment induces the expression of genes involved in healing processes in the rheumatoid arthritis synovium. Arthritis & Rheumatism. 2011 May;63(5):1246–54.

146. Akulenko R, Merl M, Helms V. BEclear: Batch Effect Detection and Adjustment in DNA Methylation Data. Deng D, editor. PLoS ONE. 2016 Aug 25;11(8):e0159921.

147. Palmer C, Diehn M, Alizadeh AA, Brown PO. Cell-type specific gene expression profiles of leukocytes in human peripheral blood. BMC Genomics. 2006 May 16;7:115.

148. mixOmics: an R package for ‘omics feature selection and multiple data integration | bioRxiv [Internet]. [cited 2020 Mar 10]. Available from: https://www.biorxiv.org/content/10.1101/108597v1

149. Jones GW, Bombardieri M, Greenhill CJ, McLeod L, Nerviani A, Rocher-Ros V, et al. Interleukin-27 inhibits ectopic lymphoid-like structure development in early inflammatory arthritis. The Journal of Experimental Medicine. 2015 Oct 19;212(11):1793–802.

150. Cancer gene expression signatures – The rise and fall? | Elsevier Enhanced Reader [Internet]. [cited 2020 May 25]. Available from: https://reader.elsevier.com/reader/sd/pii/S0959804913001536?token=959BA81B688C47713B2ED651A717621CAD3AF51315D9E79BF12B610950A754ED9D386C6569FEE1FC13F00CBFF1F01842

151. Baron D, Ramstein G, Chesneau M, Echasseriau Y, Pallier A, Paul C, et al. A common gene signature across multiple studies relate biomarkers and functional regulation in tolerance to renal allograft. Kidney Int. 2015 May;87(5):984–95.

152. Kraan TCTM van der P, Gaalen FA van, Kasperkovitz PV, Verbeet NL, Smeets TJM, Kraan MC, et al. Rheumatoid arthritis is a heterogeneous disease: Evidence for differences in the activation of the STAT-1 pathway between rheumatoid tissues. Arthritis & Rheumatism. 2003;48(8):2132–45.

153. Ungethuem U, Haeupl T, Witt H, Koczan D, Krenn V, Huber H, et al. Molecular signatures and new candidates to target the pathogenesis of rheumatoid arthritis. Physiol Genomics. 2010 Nov 29;42A(4):267–82.

154. Woetzel D, Huber R, Kupfer P, Pohlers D, Pfaff M, Driesch D, et al. Identification of rheumatoid arthritis and osteoarthritis patients by transcriptome-based rule set generation. Arthritis Research & Therapy. 2014;16(2):R84.

155. Timmer TCG, Baltus B, Vondenhoff M, Huizinga TWJ, Tak PP, Verweij CL, et al. Inflammation and ectopic lymphoid structures in rheumatoid arthritis synovial tissues dissected by genomics technology: identification of the interleukin-7 signaling pathway in tissues with lymphoid neogenesis. Arthritis Rheum. 2007 Aug;56(8):2492–502.

156. Huber R, Hummert C, Gausmann U, Pohlers D, Koczan D, Guthke R, et al. Identification of intra-group, inter-individual, and gene-specific variances in mRNA expression profiles in the rheumatoid arthritis synovial membrane. Arthritis Res Ther. 2008;10(4):R98.

157. Romão VC, Vital EM, Fonseca JE, Buch MH. Right drug, right patient, right time: aspiration or future promise for biologics in rheumatoid arthritis? Arthritis Res Ther [Internet]. 2017 [cited 2020 Mar 13];19. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5655983/

158. Singh JA, Saag KG, Bridges SL, Akl EA, Bannuru RR, Sullivan MC, et al. 2015 American College of Rheumatology Guideline for the Treatment of Rheumatoid Arthritis: ACR RA Treatment Recommendations. Arthritis Care & Research. 2016 Jan;68(1):1–25.

159. Quinn MA, Conaghan PG, Emery P. The therapeutic approach of early intervention for rheumatoid arthritis: what is the evidence? Rheumatology (Oxford). 2001 Nov 1;40(11):1211–20.

160. Burmester GR, Pope JE. Novel treatment strategies in rheumatoid arthritis. Lancet. 2017 10;389(10086):2338–48.

161. Cañete JD, Celis R, Moll C, Izquierdo E, Marsal S, Sanmartí R, et al. Clinical significance of synovial lymphoid neogenesis and its reversal after anti-tumour necrosis factor alpha therapy in rheumatoid arthritis. Ann Rheum Dis. 2009 May;68(5):751–6.

162. Daien CI, Gailhac S, Mura T, Combe B, Hahne M, Morel J. High levels of memory B cells are associated with response to a first tumor necrosis factor inhibitor in patients with rheumatoid arthritis in a longitudinal prospective study. Arthritis Res Ther. 2014;16(2):R95.

163. Daïen CI, Gailhac S, Audo R, Mura T, Hahne M, Combe B, et al. High levels of natural killer cells are associated with response to tocilizumab in patients with severe rheumatoid arthritis. Rheumatology (Oxford). 2015 Apr 1;54(4):601–8.

164. Citro A, Scrivo R, Martini H, Martire C, De Marzio P, Vestri AR, et al. CD8+ T Cells Specific to Apoptosis-Associated Antigens Predict the Response to Tumor Necrosis Factor Inhibitor Therapy in Rheumatoid Arthritis. PLoS One [Internet]. 2015 Jun 10 [cited 2020 May 19];10(6). Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4465029/

165. Chara L, Sánchez-Atrio A, Pérez A, Cuende E, Albarrán F, Turrión A, et al. The number of circulating monocytes as biomarkers of the clinical response to methotrexate in untreated patients with rheumatoid arthritis. J Transl Med [Internet]. 2015 Jan 16 [cited

2020 May 19];13. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4310181/

166. Chara L, Sánchez-Atrio A, Pérez A, Cuende E, Albarrán F, Turrión A, et al. Monocyte populations as markers of response to adalimumab plus MTX in rheumatoid arthritis. Arthritis Res Ther. 2012 Jul 27;14(4):R175.

167. R4-RA [Internet]. [cited 2020 May 21]. Available from: http://www.r4ra-nihr.whri.qmul.ac.uk/

168. Press Release: Tocilizumab More Effective than Rituximab [Internet]. [cited 2020 May 21]. Available from: https://www.rheumatology.org/About-Us/Newsroom/Press-Releases/ID/1051

169. STRAP - Stratification of Biologic Therapies for RA by Pathobiology [Internet]. [cited 2020 May 21]. Available from: http://www.matura-mrc.whri.qmul.ac.uk/aims_and_objectives.php

170. Wong KHY, Levy-Sakin M, Kwok P-Y. De novo human genome assemblies reveal spectrum of alternative haplotypes in diverse populations. Nature Communications. 2018 Aug 2;9(1):3040.

171. Li R, Tian X, Yang P, Fan Y, Li M, Zheng H, et al. Recovery of non-reference sequences missing from the human reference genome. BMC Genomics. 2019 Oct 16;20(1):746.

172. Frankish A, Diekhans M, Ferreira A-M, Johnson R, Jungreis I, Loveland J, et al. GENCODE reference annotation for the human and mouse genomes. Nucleic Acids Res. 2019 08;47(D1):D766–73.

173. Sing T, Sander O, Beerenwinkel N, Lengauer T. ROCR: visualizing classifier performance in R. Bioinformatics. 2005 Oct 15;21(20):3940–1.

174. Tak PP. Is early rheumatoid arthritis the same disease process as late rheumatoid arthritis? Best Pract Res Clin Rheumatol. 2001 Mar;15(1):17–26.

175. Smeets TJ, Dolhain RJEM null, Miltenburg AM, de Kuiper R, Breedveld FC, Tak PP. Poor expression of T cell-derived cytokines and activation and proliferation markers in early rheumatoid synovial tissue. Clin Immunol Immunopathol. 1998 Jul;88(1):84–90.

176. Genovese MC, Durez P, Richards HB, Supronik J, Dokoupilova E, Aelion JA, et al. One-year Efficacy and Safety Results of Secukinumab in Patients With Rheumatoid Arthritis: Phase II, Dose-finding, Double-blind, Randomized, Placebo-controlled Study. J Rheumatol. 2014 Mar;41(3):414–21.

177. Tahir H, Deodhar A, Genovese M, Takeuchi T, Aelion J, Van den Bosch F, et al. Secukinumab in Active Rheumatoid Arthritis after Anti-TNFα Therapy: A Randomized, Double-Blind Placebo-Controlled Phase 3 Study. Rheumatol Ther. 2017 Dec;4(2):475–88.

178. Dokoupilová E, Aelion J, Takeuchi T, Malavolta N, Sfikakis P, Wang Y, et al. Secukinumab after anti-tumour necrosis factor-α therapy: a phase III study in active rheumatoid arthritis. Scandinavian Journal of Rheumatology. 2018 Jul 4;47(4):276–81.

179. Fawcett T. An introduction to ROC analysis. Pattern Recognition Letters. 2006 Jun 1;27(8):861–74.

180. Mandrekar JN. Receiver Operating Characteristic Curve in Diagnostic Test Assessment. Journal of Thoracic Oncology. 2010 Sep 1;5(9):1315–6.

181. Emery P. Treatment of rheumatoid arthritis. BMJ. 2006 Jan 21;332(7534):152–5.

182. Firestein GS. Pathogenesis of rheumatoid arthritis: how early is early? Arthritis Research & Therapy. 2005 Jun 17;7(4):157.

183. Niemantsverdriet E, Dougados M, Combe B, Mil AHM van der H. Referring early arthritis patients within 6 weeks versus 12 weeks after symptom onset: an observational cohort study. The Lancet Rheumatology. 2020 Jun 1;2(6):e332–8.

184. De Cock D, Van der Elst K, Stouten V, Peerboom D, Joly J, Westhovens R, et al. The perspective of patients with early rheumatoid arthritis on the journey from symptom onset until referral to a rheumatologist. Rheumatol Advanc Pract [Internet]. 2019 Jul 1 [cited 2020 Oct 2];3(2). Available from: https://academic.oup.com/rheumap/article/3/2/rkz035/5556820

185. Scott DL. Early rheumatoid arthritis. Br Med Bull. 2007 Jan 1;81–82(1):97–114.

186. Is early rheumatoid arthritis the same disease process as late rheumatoid arthritis? | Elsevier Enhanced Reader [Internet]. [cited 2020 Jun 11]. Available from: https://reader.elsevier.com/reader/sd/pii/S1521694200901232?token=DB1CAE7FA6290A195CC07F0EADF3C7371B807938BBA8CE36226B2B849655F54CFA66FA4D2FE59113FFDECE0AE60F281C

187. Kraan MC, Haringman JJ, Post WJ, Versendaal J, Breedveld FC, Tak PP. Immunohistological analysis of synovial tissue for differential diagnosis in early arthritis. Rheumatology (Oxford). 1999 Nov 1;38(11):1074–80.

188. Katrib A, Tak PP, Bertouch JV, Cuello C, McNeil HP, Smeets TJM, et al. Expression of chemokines and matrix metalloproteinases in early rheumatoid arthritis. Rheumatology. 2001 Sep;40(9):988–94.

189. Smeets TJM, Dolhain RJEM, Miltenburg AMM, de Kuiper R, Breedveld FC, Tak PP. Poor Expression of T Cell-Derived Cytokines and Activation and Proliferation Markers in Early Rheumatoid Synovial Tissue. Clinical Immunology and Immunopathology. 1998 Jul;88(1):84–90.

190. Tak PP, Smeets TJM, Daha MR, Kluin PM, Meijers KAE, Brand R, et al. Analysis of the synovial cell infiltrate in early rheumatoid synovial tissue in relation to local disease activity. Arthritis & Rheumatism. 1997;40(2):217–25.

191. Raza K. Early rheumatoid arthritis is characterised by a distinct and transient synovial fluid cytokine profile of T cell and stromal cell origin. Arthritis Res Ther. 2019 Dec;21(1):226, s13075-019-2026–4.

192. Chen Y-F, Jobanputra P, Barton P, Jowett S, Bryan S, Clark W, et al. A systematic review of the effectiveness of adalimumab, etanercept and infliximab for the treatment of rheumatoid arthritis in adults and an economic evaluation of their cost-effectiveness. Health Technol Assess. 2006 Nov;10(42):iii–iv, xi–xiii, 1–229.

193. Wessels J a. M, Huizinga TWJ, Guchelaar H-J. Recent insights in the pharmacological actions of methotrexate in the treatment of rheumatoid arthritis. Rheumatology (Oxford, England). 2008 Mar;47(3):249–55.

194. Goodsell DS. The Molecular Perspective: Methotrexate. The Oncologist. 1999;4(4):340–1.

195. Anolik JH, Ravikumar R, Barnard J, Owen T, Almudevar A, Milner ECB, et al. Cutting Edge: Anti-Tumor Necrosis Factor Therapy in Rheumatoid Arthritis Inhibits Memory B Lymphocytes via Effects on Lymphoid Germinal Centers and Follicular Dendritic Cell Networks. The Journal of Immunology. 2008 Jan 15;180(2):688–92.

196. Walsh AM, Wechalekar MD, Guo Y, Yin X, Weedon H, Proudman SM, et al. Triple DMARD treatment in early rheumatoid arthritis modulates synovial T cell activation and plasmablast/plasma cell differentiation pathways. Kuwana M, editor. PLOS ONE. 2017 Sep 1;12(9):e0183928.

197. Rivellese F, Humby F, Bugatti S, Fossati-Jimack L, Rizvi H, Lucchesi D, et al. B Cell Synovitis and Clinical Phenotypes in Rheumatoid Arthritis: Relationship to Disease Stages and Drug Exposure. Arthritis Rheumatol. 2020 May;72(5):714–25.

198. Brennan FM, McInnes IB. Evidence that cytokines play a role in rheumatoid arthritis. The Journal of Clinical Investigation. 2008 Nov 3;118(11):3537.

199. Nowell MA, Richards PJ, Horiuchi S, Yamamoto N, Rose-John S, Topley N, et al. Soluble IL-6 Receptor Governs IL-6 Activity in Experimental Arthritis: Blockade of Arthritis Severity by Soluble Glycoprotein 130. J Immunol. 2003 Sep 15;171(6):3202–9.

200. Alivernini S, Tolusso B, Petricca L, Bui L, Di Mario C, Gigante MR, et al. Synovial Predictors of Differentiation to Definite Arthritis in Patients With Seronegative Undifferentiated Peripheral Inflammatory Arthritis: microRNA Signature, Histological, and Ultrasound Features. Front Med [Internet]. 2018 [cited 2020 Mar 10];5. Available from: https://www.frontiersin.org/articles/10.3389/fmed.2018.00186/full

201. Lee C, Oh J-I, Park J, Choi J-H, Bae E-K, Lee HJ, et al. TNFα Mediated IL-6 Secretion Is Regulated by JAK/STAT Pathway but Not by MEK Phosphorylation and AKT Phosphorylation in U266 Multiple Myeloma Cells. Biomed Res Int [Internet]. 2013 [cited 2020 Oct 18];2013. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3787550/

202. Nowell MA, Williams AS, Carty SA, Scheller J, Hayes AJ, Jones GW, et al. Therapeutic Targeting of IL-6 Trans Signaling Counteracts STAT3 Control of Experimental Inflammatory Arthritis. The Journal of Immunology. 2009 Jan 1;182(1):613–22.

203. Isomäki P, Junttila I, Vidqvist K-L, Korpela M, Silvennoinen O. The activity of JAK-STAT pathways in rheumatoid arthritis: constitutive activation of STAT3 correlates with interleukin 6 levels. Rheumatology. 2015 Jan 6;54(6):1103–13.

204. Nielsen MA, Lomholt S, Mellemkjær A, Andersen MN, Buckley CD, Kragstrup TW. Responses to Cytokine Inhibitors Associated with Cellular Composition in Models of Immune-Mediated Inflammatory Arthritis. ACR Open Rheumatology. 2020;2(1):3–10.

205. Kaur K, Kalra S, Kaushal S. Systematic Review of Tofacitinib: A New Drug for the Management of Rheumatoid Arthritis. Clinical Therapeutics. 2014 Jul 1;36(7):1074–86.

206. Gordon RA, Grigoriev G, Lee A, Kalliolias GD, Ivashkiv LB. The IFN signature and STAT1 expression in RA synovial fluid macrophages are induced by TNFα and counter-regulated by synovial fluid microenvironment. Arthritis Rheum. 2012 Oct;64(10):3119–28.

207. Hooge D, K AS, Loo VD, J FA, Koenders MI, Bennink MB, et al. Local activation of STAT-1 and STAT-3 in the inflamed synovium during zymosan-induced arthritis: Exacerbation of joint inflammation in STAT-1 gene–knockout mice. Arthritis & Rheumatism. 2004 Jun 1;50(6):2014–23.

208. Hirahara K, Onodera A, Villarino AV, Bonelli M, Sciumè G, Laurence A, et al. Asymmetric Action of STAT Transcription Factors Drives Transcriptional Outputs and Cytokine Specificity. Immunity. 2015 May 19;42(5):877–89.

209. Fielding CA, McLoughlin RM, McLeod L, Colmont CS, Najdovska M, Grail D, et al. IL-6 Regulates Neutrophil Trafficking during Acute Inflammation via STAT3. J Immunol. 2008 Aug 1;181(3):2189–95.

210. Hückel M, Schurigt U, Wagner AH, Stöckigt R, Petrow PK, Thoss K, et al. Attenuation of murine antigen-induced arthritis by treatment with a decoy oligodeoxynucleotide inhibiting signal transducer and activator of transcription-1 (STAT-1). Arthritis Research & Therapy. 2005 Dec 30;8(1):R17.

211. Wang S, Wang L, Wu C, Sun S, Pan J. E2F2 directly regulates the STAT1 and PI3K/AKT/NF-κB pathways to exacerbate the inflammatory phenotype in rheumatoid arthritis synovial fibroblasts and mouse embryonic fibroblasts. Arthritis Research & Therapy. 2018 Oct 4;20(1):225.

212. Adamson AS, Collins K, Laurence A, O'Shea JJ. The Current STATus of lymphocyte signaling: new roles for old players (STATs in lymphocyte signaling). Current opinion in immunology. 2009 Apr;21(2):161.

213. Kasperkovitz P, Verbeet N, Smeets T, van Rietschoten JGI, Kraan M, van der Pouw Kraa. . TCTM, et al. Activation of the STAT1 pathway in rheumatoid arthritis. Ann Rheum Dis. 2004 Mar;63(3):233–9.

214. Avalle L, Pensa S, Regis G, Novelli F, Poli V. STAT1 and STAT3 in tumorigenesis. JAKSTAT. 2012 Apr 1;1(2):65–72.

215. Wan C-K, Andraski AB, Spolski R, Li P, Kazemian M, Oh J, et al. Opposing roles of STAT1 and STAT3 in IL-21 function in CD4+ T cells. PNAS. 2015 Jul 28;112(30):9394–9.

216. Wang Y, Boxel-Dezaire AHH van, Cheon H, Yang J, Stark GR. STAT3 activation in response to IL-6 is prolonged by the binding of IL-6 receptor to EGF receptor. Proceedings of the National Academy of Sciences of the United States of America. 2013 Oct 15;110(42):16975.

217. Krämer A, Green J, Pollard J, Tugendreich S. Causal analysis approaches in Ingenuity Pathway Analysis. Bioinformatics. 2014 Feb 15;30(4):523–30.

218. Rouillard AD, Gundersen GW, Fernandez NF, Wang Z, Monteiro CD, McDermott MG, et al. The harmonizome: a collection of processed datasets gathered to serve and mine knowledge about genes and proteins. Database (Oxford) [Internet]. 2016 Jan 1 [cited 2020 Oct 18];2016. Available from: https://academic.oup.com/database/article/doi/10.1093/database/baw100/2630482

219. Mi H, Muruganujan A, Thomas PD. PANTHER in 2013: modeling the evolution of gene function, and other gene attributes, in the context of phylogenetic trees. Nucleic Acids Res. 2013 Jan;41(Database issue):D377-386.

220. ENCODE Project Consortium, Moore JE, Purcaro MJ, Pratt HE, Epstein CB, Shoresh N, et al. Expanded encyclopaedias of DNA elements in the human and mouse genomes. Nature. 2020 Jul;583(7818):699–710.

221. Satoh J, Tabunoki H. A Comprehensive Profile of ChIP-Seq-Based STAT1 Target Genes Suggests the Complexity of STAT1-Mediated Gene Regulatory Mechanisms. Gene Regulation and Systems Biology. 2013;7:41.

222. Bhinge AA, Kim J, Euskirchen GM, Snyder M, Iyer VR. Mapping the chromosomal targets of STAT1 by Sequence Tag Analysis of Genomic Enrichment (STAGE). Genome Res. 2007 Jan 6;17(6):910–6.

223. Hutchins AP, Poulain S, Miranda-Saavedra D. Genome-wide analysis of STAT3 binding in vivo predicts effectors of the anti-inflammatory response in macrophages. Blood. 2012 Mar 29;119(13):e110–9.

224. Robinson JT, Thorvaldsdóttir H, Winckler W, Guttman M, Lander ES, Getz G, et al. Integrative genomics viewer. Nature Biotechnology. 2011 Jan;29(1):24–6.

225. Minimum information about a microarray experiment (MIAME)—toward standards for microarray data | Nature Genetics [Internet]. [cited 2020 Nov 3]. Available from: https://www.nature.com/articles/ng1201-365

226. Wilkinson MD, Dumontier M, Aalbersberg IjJ, Appleton G, Axton M, Baak A, et al. The FAIR Guiding Principles for scientific data management and stewardship. Sci Data [Internet]. 2016 Mar 15 [cited 2020 Nov 5];3. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4792175/