Epigenetics of response to biologic drug therapy in rheumatoid arthritis

A thesis submitted to the University of Manchester for the degree of Doctor of Philosophy in the Faculty of Medical and Human Sciences

2014

Amy Philomena Webster

Arthritis Research UK Centre for Genetics and Genomics
Centre for Musculoskeletal Research
Institute of Inflammation and Repair
School of Medicine
# Table of Contents

1. Introduction ........................................................................................................................................... 29

1.1. Rheumatoid arthritis .......................................................................................................................... 30

1.1.1. Epidemiology ................................................................................................................................... 30

1.1.2. Clinical features .............................................................................................................................. 32

1.1.3. Pathogenesis ................................................................................................................................... 34

1.1.4. Comorbidity .................................................................................................................................... 36

1.1.5. Economic cost of rheumatoid arthritis ......................................................................................... 36

1.2. Treatment of rheumatoid arthritis .................................................................................................... 37

1.2.1. DMARDs ......................................................................................................................................... 37

1.2.2. Biologics .......................................................................................................................................... 38

1.2.3. Anti-TNF therapies ....................................................................................................................... 40

1.2.4. Future treatment of rheumatoid arthritis ....................................................................................... 45

1.2.5. Key research questions .................................................................................................................. 46

1.2.6. Non-genetic risk factors in the development of rheumatoid arthritis ........................................... 47

1.2.7. Genetic risk factors for susceptibility to rheumatoid arthritis ..................................................... 49

1.3. Clinical predictors of response to anti-TNF therapy ......................................................................... 57

1.3.1. Importance of predictors of response to anti-TNF therapy ......................................................... 57

1.3.2. Classifying treatment response ..................................................................................................... 58

1.3.3. Gender as an indicator of response to anti-TNF therapy .............................................................. 58

1.3.4. Initial disability as an indicator of response to anti-TNF therapy ................................................. 60

1.3.5. Dosage effects and adherence in anti-TNF response .................................................................... 60
Concurrent use of DMARDs as an indicator of response to anti-TNF therapy .................................................................61

Presence of autoantibodies as an indicator of response to anti-TNF therapy ...........................................................................62

The influence of environmental factors on response to anti-TNF therapy .....................................................................................63

Serum and gene expression biomarkers .................................................63

Genetic factors as indicators of response to anti-TNF therapy .................64

Epigenetics .................................................................................................................73

Histone modifications ............................................................................................74

Non-coding RNA .....................................................................................................75

DNA methylation ....................................................................................................76

Heritability of DNA methylation ..........................................................................83

DNA methylation research techniques ..............................................................85

Considerations in DNA methylation study design ............................................92

DNA methylation in rheumatoid arthritis .......................................................96

DNA methylation in other inflammatory diseases .........................................101

Epigenetic influences in therapy response .......................................................106

Summary ..................................................................................................................109

Aims .........................................................................................................................110

Methods .................................................................................................................112

Sub-study sample selection ....................................................................................113

Etanercept and adalimumab response study .................................................113

Measuring treatment response ..........................................................................114
2.1.3. Rheumatoid arthritis discordant monozygotic twins study ..........114
2.1.4. Investigation into the effect of cryopreservation on methylation in
CD3+CD4+ T-cells .................................................................115
2.2. Techniques .............................................................................116
2.2.1. Isolation of cryopreserved and fresh CD3+CD4+ T-cells ..........116
2.2.2. DNA extraction .................................................................120
2.2.3. Quality control of DNA .......................................................122
2.2.4. Bisulphite conversion .........................................................124
2.2.5. Epigenome-wide association studies ...................................127
2.2.6. Pyrosequencing .................................................................137
2.3. Statistical analysis .................................................................149
2.3.1. Epigenome-wide association studies of anti-TNF response ....149
2.3.2. Epigenome-wide association study of rheumatoid arthritis
discordant monozygotic twins .....................................................161
2.3.3. Investigation into effects of cryopreservation of cells on DNA
methylation .............................................................................166
3. Results ......................................................................................169
3.1. Quality control of bisulphite conversion and validation of pyrosequencing
technology with Sanger sequencing ..............................................170
3.1.1. Sanger sequencing of DAPK1 ..............................................170
3.1.2. Pyrosequencing of DAPK1_3 and DAPK1_5 ........................174
3.1.3. Pyrosequencing of DAPK1_1 ..............................................177
3.1.4. Summary of bisulphite conversion quality control experiment
results ......................................................................................180
3.2. Epigenome-wide association studies of anti-TNF treatment response 181

3.2.1. Epigenome-wide association study of etanercept response .................. 181

3.2.2. Epigenome-wide association study of adalimumab response .......... 210

3.2.3. Combined analysis of epigenome-wide association study of etanercept and adalimumab response ........................................ 230

3.3. Epigenome-wide association study of rheumatoid arthritis discordant monozygotic twins ................................................................. 236

3.3.1. Patient characteristics ................................................................. 236

3.3.2. Analysis of whole blood twin pairs ............................................. 238

3.3.3. Validation using rheumatoid arthritis discordant twin PBMC samples ........................................................................................................ 248

3.3.4. Summary of results of the investigation into methylation in rheumatoid arthritis discordant monozygotic twins .............................. 257

3.4. Investigation into the effect of cryopreservation on DNA methylation in CD3+CD4+ T-cells ........................................................................................................ 258

3.4.1. Patient characteristics ................................................................. 258

3.4.2. Preparation of CD3+CD4+ T-cells by FACS ................................. 259

3.4.3. Whole genome DNA methylation profiling of CD3+CD4+ T-cells ..... 262

3.4.4. Summary of cryopreservation experiment results ..................... 274

4. Discussion .......................................................................................... 275

4.1. Overview of findings ........................................................................ 276

4.2. Strengths and weaknesses of studies ............................................. 277

4.2.1. Establishing techniques in the lab ............................................. 277

4.2.2. Methylation in anti-TNF response ............................................. 278
4.2.3. Methylation in rheumatoid arthritis discordant monozygotic twins ......284
4.2.4. Investigation into effect of cryopreservation on methylation ..........287
4.3. Meaning of studies: possible mechanisms and implications..............289
4.3.1. Anti-TNF response studies ..............................................................289
4.3.2. Rheumatoid arthritis discordant monozygotic twin study...............297
4.3.3. Effect of cryopreservation on methylation .......................................301
4.4. Future work .........................................................................................302
4.4.1. Replication of response to anti-TNF methylation differences .........302
4.4.2. Investigation into alternative predictors of anti-TNF treatment response............................................................................................................305
4.4.3. Investigation into the mechanism by which LRPAP1 methylation could be affecting anti-TNF response ..............................................................307
4.4.4. Establishing the temporal origins of rheumatoid arthritis methylation changes.................................................................................................308
4.4.5. Characterising the functional effects of methylation changes in rheumatoid arthritis.................................................................309
4.5. Conclusion...............................................................................................310

Total word count: 70,870
List of Tables

Table 1.1: Summary of the 2010 classification criteria for RA from the American College of Rheumatology/European League Against Rheumatism (EULAR). ................................................................. 33

Table 1.2: TNF Inhibitors used in the treatment of rheumatoid arthritis (adapted from (Scott 2012; Wilkie and Schwieterman 2012). ......................... 41

Table 1.3: Summary of the results of studies of TNF-308 promoter genotypes and their affect on response to specific anti-TNF drugs (adapted from (Emery and Dorner 2011). ......................................................... 66

Table 1.4: Techniques which employ different strategies for methylation analysis of DNA (adapted from (Laird 2010)). ........................................ 86

Table 2.1: Order, name and quantity of reagents added to Flow-Through chamber in initial stages of single-base extension during BeadChip preparation. ........................................................................ 134

Table 2.2: Order of addition, incubation periods and repeats for each reagent in staining protocol during preparation of BeadChips. ................................. 135

Table 2.3: Primer sequences from Human Methylated & Non-Methylated DNA set from Zymo Research, used in Sanger sequencing of bisulphite converted DNA. ................................................................. 145

Table 2.4: Thermocycler conditions used for the amplification of bisulphite converted DNA using DAPK1 primers in preparation for Sanger sequencing. ........................................................................ 145

Table 2.5: Sequences of forward and reverse sequencing primers for the sequencing of DAPK1 using Sanger sequencing .............................................. 146

Table 2.6: Primer sequences used in pyrosequencing of two regions of the DAPK1 gene, DAPK1_3 and DAPK1_5. ......................................................... 147

Table 2.7: Expected intensities of control probes. .................................................. 152
Table 3.1: Patient characteristics for epigenome-wide association study of etanercept response. ................................................................. 181

Table 3.2: Amount of variance in the data explained by each principal component loading. ................................................................. 188

Table 3.3: Correlation of first four principal component loadings with batch and patient specific measurements. .............................................. 189

Table 3.4: Differentially methylated positions between responders and non-responders to etanercept.............................................................. 195

Table 3.5: Differentially methylated positions which passed false discovery rate of 0.05 in etanercept response analysis in which SNP-containing probes were included............................................................... 198

Table 3.6: Cell composition estimations for responders and non-responders to etanercept. .............................................................................. 200

Table 3.7: Comparison of differentially methylated positions associated with response to etanercept in data adjusted for cell composition versus data not adjusted for cell composition........................................... 203

Table 3.8: Power calculation for EWAS of etanercept response at a significance level of p=0.05 ........................................................................... 205

Table 3.9: Pyrosequencing assays for validation of top 5 most differentially methylated positions from etanercept response epigenome-wide association study. ................................................................. 207

Table 3.10: Correlation of HumanMethylation450 BeadChip results with pyrosequencing results from the top most differentially methylated positions. ..................................................................................... 208

Table 3.11: Patient characteristics for epigenome-wide association study of adalimumab response. .............................................................. 210

Table 3.12: Correlation of first four PCs with batch and patient characteristics in analysis of methylation in adalimumab treatment response. .............. 216
Table 3.13: Top 20 CpG sites at which methylation is associated with response to adalimumab (not adjusted for cell composition).................................219

Table 3.14: Proportions of cells in responders and non-responders to adalimumab.............................................................................................................220

Table 3.15: Top 20 most differentially methylated positions between responders and non-responders to adalimumab with and without adjustment for cell composition..........................................................223

Table 3.16: Pyrosequencing assays for validation of top 5 most differentially methylated positions from adalimumab response epigenome-wide association study........................................................................225

Table 3.17: Correlation of pyrosequencing with HumanMethylation450 BeadChip results for differentially methylated positions in Adalimumab response.........................................................................................227

Table 3.18: Summary of number of patients included in combined analysis of etanercept and adalimumab data...............................................................230

Table 3.19: Associations between the first four principal components and possible batch effects or patient characteristics...........................................232

Table 3.20: Summary of RA discordant twin pairs.................................................................................................................................237

Table 3.21: Cohort characteristics for RA discordant twins with whole blood samples............................................................................................237

Table 3.22: Contribution of each principal component (PC) to variability of dataset.........................................................................................241

Table 3.23: P-values for the correlation of the first four PC loadings with possible batch effects and patient characteristics in whole blood RA discordant monozygotic twin methylation analysis........................................241

Table 3.24: Differentially methylated positions (DMPs) between rheumatoid arthritis discordant monozygotic twins identified in whole blood samples, unadjusted for cell composition.................................244
Table 3.25: Cell composition estimates calculated from whole blood HumanMethylation450 BeadChip data from RA discordant MZ twins. ................................................................. 245

Table 3.26: Differentially methylated positions in whole blood samples from RA discordant MZ twins, after adjustment for cell composition. ............. 247

Table 3.27: P-values of association between the first four principal component loadings and potential batch effects. .................................................. 250

Table 3.28: Differentially methylated positions identified in PBMC samples from RA discordant MZ twins. ............................................................... 252

Table 3.29: Cell proportion estimates in RA and non-RA twins. ......................... 253

Table 3.30: Differentially methylated positions identified in PBMC samples from RA discordant MZ twins with adjustment for cell composition. ........ 256

Table 3.31: Age and gender of healthy volunteers. ........................................... 258

Table 3.32: Counts of fresh PBMCs, and CD3+CD4+ T-cell counts from fresh and cryopreserved samples obtained from 10ml blood. ................. 261

Table 3.33: Contribution of each of the first ten principal components to variability in dataset of cryopreserved and fresh CD3+CD4+ T cells... 266

Table 3.34: Correlation of each of the first four principal components with possible batches and volunteer characteristics in the analysis of cryopreserved and fresh CD3+CD4+ T cells........................................... 267

Table 3.35: Most differentially methylated positions between cryopreserved and fresh samples of CD3+CD4+ T-cells............................................. 269

Table 3.36: Pearson’s correlation for methylation values in cryopreserved versus fresh samples from each individual in the investigation of effect of cryopreservation on methylation............................................. 270
List of Figures

Figure 1.1: Incidence of rheumatoid arthritis according to age and gender (Uhlig et al. 1998). ..............................................................31

Figure 1.2: Rheumatoid arthritis affected joint compared to healthy joint, showing inflamed synovium......................................................35

Figure 1.3: Interaction between different classes of biologic drugs and the inflammatory pathway in rheumatoid arthritis.................................39

Figure 1.4: Main adverse effects associated with use of anti-TNF therapy (Willrich, Murray, and Snyder 2014)............................................42

Figure 1.5: Rheumatoid arthritis genetic risk loci in order of discovery. ...................55

Figure 1.6: Simplistic postulated mechanism by which DNA methylation in a promoter can influence transcription of a gene. .........................79

Figure 1.7: Methylation levels in regions relative to a gene have expression-dependent profiles..........................................................80

Figure 1.8: Chemical structure of modified and unmodified cytosine nucleotides. …82

Figure 1.9: Heat map showing methylation of 49 genes identified as differentially methylated in 5 systemic lupus erythematosus discordant twin pairs and matched controls. ..............................102

Figure 2.1: Prepared leucosep tubes before and after centrifugation during PBMC isolation. .................................................................117

Figure 2.2: Example of good and poor quality DNA sample, assessed using gel electrophoresis. .............................................................122

Figure 2.3: Mode of action of sodium bisulphite (HSO3- ) in the bisulphite conversion of DNA...............................................................124

Figure 2.4: Example of translation of epigenetic signature into a genetic signature during bisulphite conversion of DNA and subsequent PCR........................................................125
Figure 2.5: Summary of Infinium I and Infinium II chemistries present on the Illumina HumanMethylation450 BeadChip. ......................................................... 128
Figure 2.6: Photo of the Illumina HumanMethylation450 BeadChip. ................. 129
Figure 2.7: Flow-through chamber assembly of BeadChip used during single-base extension and staining steps of BeadChip preparation................. 133
Figure 2.8: Principles of pyrosequencing technology. ........................................ 138
Figure 2.9: Setup of the vacuum workstation used to create single stranded amplicons for use in pyrosequencing...................................................... 141
Figure 2.10: Location of PyroMarks Gold Q24 reagents in the cartridge. .......... 143
Figure 2.11: Summary of methylation analysis pipeline......................................... 150
Figure 2.12: Example of distribution of beta values of a sample before and after SWAN. ............................................................................................................. 156
Figure 2.13: Summary of analysis pipeline for identification of methylation differences in RA discordant monozygotic twins................................. 162
Figure 3.1: Sanger sequencing traces for 24bp section of DAPK1....................... 172
Figure 3.2: Sanger sequencing trace showing long mononucleotide run causing sequence slippage and subsequent mixed signal. ......................... 173
Figure 3.3: Example pyrosequencing trace of DAPK1_5 assay of a biological DNA sample. ............................................................................................................. 176
Figure 3.4: Pyrogram indicating higher than expected peak height in assay. ...... 176
Figure 3.5: Pyrosequencing results from DAPK1_1 assay of methylated control DNA..................................................................................................................... 178
Figure 3.6: Example of pyrosequencing results from DAPK1_1 assay of a biological DNA sample.............................................................. 179
Figure 3.7: Example of control dashboard produced in GenomeStudio for bisulphite conversion control probes targeting type 2 probes.......... 183
Figure 3.8: Quality control kernel density plot produced in minfi showing distribution of beta values split into good and poor responders to etanercept. ................................................................. 184

Figure 3.9: Quality control graph produced using quality control information extracted from the unnormalised methylset object in minfi. ............... 185

Figure 3.10: Graph indicating the variability in the dataset which is explained by each principal component................................. 188

Figure 3.11: Relationship between position on array of sample and the first principal component in the data, prior to ComBat adjustment. .......... 191

Figure 3.12: Relationship between position of sample on array and the first principal component after ComBat adjustment. ...................... 191

Figure 3.13: Manhattan plot of the relationship between response to etanercept and DNA methylation..................................................... 194

Figure 3.14: Position of CpG island in relation to the LRPAP1 gene and position of the differentially methylated CpG sites.......................... 195

Figure 3.15: Manhattan Plot of analysis of etanercept EWAS response results, with SNP-containing probes included. .................................. 197

Figure 3.16: Position of CpG island in relation to LRPAP1 gene and position of differentially methylated CpG sites across island, including previously excluded SNP-containing probes.......................... 199

Figure 3.17: Manhattan Plot of etanercept response EWAS data after adjustment for cell composition .............................................. 204

Figure 3.18: Correlation of the percentage methylation between HumanMethylation450 BeadChip (x-axis) and pyrosequencing data for the most differentially methylated position, cg04857395............ 208

Figure 3.19: Example of GenomeStudio Control plots used to assess performance of arrays and identify sample-dependent or sample-independent problems. .................................................. 211
Figure 3.20: Density plot of beta values from each sample coloured by good and poor response to adalimumab. .................................................................212

Figure 3.21: Quality control graph produced using QC information extracted from the unnormalised methylset object in minfi. .................................213

Figure 3.22: Variability of methylation data explained by each PC in the adalimumab treatment response study. ......................................................215

Figure 3.23: Manhattan plot of association between CpG sites and response to adalimumab. ................................................................................218

Figure 3.24: Manhattan plot of association between CpG sites and response to adalimumab with adjustment for cell composition. .........................221

Figure 3.25: Example of a program showing percentage methylation at each CpG site (highlighted within the blue bars) measured by the proportion of peak signal. .................................................................226

Figure 3.26: Correlation of methylation values of four CpG sites measured by pyrosequencing (y-axis) versus HumanMethylation450 BeadChip signal (x-axis) in validation study. ......................................................226

Figure 3.27: Correlation of cg09044186 DNA methylation levels as measured by pyrosequencing (y-axis) versus HumanMethylation450 BeadChip signal (x-axis). .................................................................................227

Figure 3.28: Correlation of cg05857826 DNA methylation levels as measured by pyrosequencing (y-axis) versus HumanMethylation450 BeadChip signal (x-axis). .................................................................................228

Figure 3.29: Manhattan plot of association between methylation levels and response to anti-TNF therapies. ..........................................................234

Figure 3.30: Differentially methylated positions in the overlapping CRYZ/TYW3 gene regions associated with response to anti-TNF therapies.........235

Figure 3.31: Quality control kernel density plot of beta value distribution in RA discordant MZ twin samples. .........................................................238
Figure 3.32: Quality control plot of log transformed median methylated versus unmethylated intensities for RA discordant twin samples..................239

Figure 3.33: Graph indicating the contribution of each principal component loading to the variability in the dataset of RA discordant MZ twins.…..240

Figure 3.34: Manhattan plot of differentially methylated positions in RA discordant MZ twin pairs in whole blood samples, unadjusted for cell composition. .................................................................243

Figure 3.35: Manhattan plot of differentially methylated positions between rheumatoid arthritis discordant monozygotic twins after adjustment for cell composition, identified in whole blood. .........................246

Figure 3.36: Contribution of principal components to variability in the HumanMethylation450 dataset of rheumatoid arthritis discordant monozygotic twins in PBMC samples. ..................................................249

Figure 3.37: Manhattan Plot for association between differentially methylated positions and RA status of MZ twins in PBMC samples. .................251

Figure 3.38: Manhattan plot of differentially methylated positions identified in PBMC samples from RA discordant MZ twins with adjustment for cell composition.................................................................255

Figure 3.39: Forward and Side scatter plot of PBMC sample before and after sorting, showing gate selection (P2) .............................................260

Figure 3.40: Forward and side scatter plots of post-sort CD3+CD4+ T-cells from the same individual, performed on fresh PBMCs (left) and cryopreserved PBMCs (right)...............................................................260

Figure 3.41: Quality control kernal density plot showing distribution of beta values of methylation for each sample in the investigation of the effect of cryopreservation on methylation.................................................262
Figure 3.42: Quality control plot of log transformed median methylated and unmethylated signal for each sample in the investigation of the effect of cryopreservation on methylation..............................263

Figure 3.43: Multiple density scaling plot of the distribution of the 1000 most variable positions in samples in the investigation of the effect of cryopreservation on methylation. ..................................................264

Figure 3.44: Principal components and the contribution of each loading to variability in methylation profile dataset of CD3+CD4+ T cells which had been prepared from cryopreserved or fresh samples.............266

Figure 3.45: Manhattan plot of differentially methylated positions between cryopreserved and fresh samples of CD3+CD4+ T-cells. ............268

Figure 3.46: Cigar plot of correlation of methylation values in cryopreserved (y-axis) versus fresh (x-axis) samples from one individual (NRHV006)...271

Figure 3.47: Cigar plot of methylation values in PBMC sample passed through FACS (x-axis) versus a PBMC sample from the same individual that wasn't exposed to FACS (y-axis)..........................................................273

Figure 4.1: Proposed mechanism by which differential DNA methylation in the LRPAP1 gene body may affect response to etanercept in patients with rheumatoid arthritis.................................................................291
Abstract

**Background:** Rheumatoid arthritis (RA) is a common complex autoimmune disorder which is influenced by both genetic and environmental factors. While multiple factors that influence susceptibility to and outcome of disease have been identified there is still a large proportion of missing heritability and limited understanding of disease pathogenesis. In recent years, biologic drug therapies have advanced treatment of RA; however good disease control is achieved in just 30% of patients, making identification of predictors of treatment response important. One area of research which is yet to be explored in relation to treatment response, and requires further evaluation in RA susceptibility, is epigenetics. Epigenetics is the study of modifications of the DNA which can influence gene expression but do not alter genetic sequence, and the most commonly studied epigenetic phenomenon, to date, is DNA methylation.

**Objectives:** To identify DNA methylation signatures predictive of treatment response to anti-TNF biologics, to explore the role of DNA methylation in RA susceptibility using disease discordant monozygotic (MZ) twins, and to assess the effect of cryopreservation of cells on DNA methylation.

**Methods:** Genome-wide DNA methylation levels were measured using the HumanMethylation450 BeadChip in pre-treatment whole blood DNA samples from individuals who had extremely good or extremely poor response to the anti-TNF therapies, etanercept and adalimumab, and in MZ twins discordant for RA (n=79 pairs). I also compared genome-wide methylation in cells which had been cryopreserved with fresh cells, to investigate if this technique is suitable for epigenetic investigations.

**Results:** I identified four methylation sites which were significantly related to response to etanercept at a false discovery rate of 5%, the most significantly
differentially methylated of which maps to the LRPAP1 gene ($p=1.46\times10^{-8}$). Indeed, four other sites at the same locus also showed evidence for differential methylation indicating that this represents a differentially methylated region. No sites were significantly associated with response to adalimumab after correction for multiple testing. I identified subtle differences in DNA methylation between RA discordant twins. Although these were not statistically significant following adjustment for cell composition, one of the most differentially methylated positions mapped to the ZNF74 gene ($p=4.97\times10^{-6}$), and replicated a methylation difference identified in the largest previous epigenome-wide association study of RA cases and unrelated healthy controls. I found that cryopreservation of cells does not significantly alter the methylome, an important observation that will impact upon design of future studies.

**Conclusions:** In the largest studies of DNA methylation in RA treatment response and RA discordant MZ twins to date, I identified significant differential methylation in etanercept response, but not adalimumab response, and found small differences in methylation in RA discordant MZ twins. I also concluded that cryopreservation does not significantly alter the methylome.
Declaration

I declare that no portion of the work referred to in the thesis has been submitted in support of an application for another degree or qualification of this or any other university or other institute of learning.
Copyright

i. The author of this thesis (including any appendices and/or schedules to this thesis) owns certain copyright or related rights in it (the “Copyright”) and s/he has given The University of Manchester certain rights to use such Copyright, including for administrative purposes.

ii. Copies of this thesis, either in full or in extracts and whether in hard or electronic copy, may be made only in accordance with the Copyright, Designs and Patents Act 1988 (as amended) and regulations issued under it or, where appropriate, in accordance with licensing agreements which the University has from time to time. This page must form part of any such copies made.

iii. The ownership of certain Copyright, patents, designs, trade marks and other intellectual property (the “Intellectual Property”) and any reproductions of copyright works in the thesis, for example graphs and tables (“Reproductions”), which may be described in this thesis, may not be owned by the author and may be owned by third parties. Such Intellectual Property and Reproductions cannot and must not be made available for use without the prior written permission of the owner(s) of the relevant Intellectual Property and/or Reproductions.

iv. Further information on the conditions under which disclosure, publication and commercialisation of this thesis, the Copyright and any Intellectual Property and/or Reproductions described in it may take place is available in the University IP Policy (see http://www.campus.manchester.ac.uk/medialibrary/policies/intellectual-property.pdf), in any relevant Thesis restriction declarations deposited in the University Library, The University Library’s regulations (see http://www.manchester.ac.uk/library/aboutus/regulations) and in The University’s policy on presentation of Theses.
## List of abbreviations

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>ACPA</td>
<td>Anti-citrullinated protein antibody</td>
</tr>
<tr>
<td>ACR</td>
<td>American college of rheumatology</td>
</tr>
<tr>
<td>5hmC</td>
<td>5-hydroxymethylcytosine</td>
</tr>
<tr>
<td>APS</td>
<td>Adenosine 5' phosphosulfate</td>
</tr>
<tr>
<td>ARUK</td>
<td>Arthritis Research UK</td>
</tr>
<tr>
<td>ATM</td>
<td>Anti-stain two-colour master mix</td>
</tr>
<tr>
<td>ATP</td>
<td>Adenosine triphosphate</td>
</tr>
<tr>
<td>BLAST</td>
<td>Basic local alignment search tool</td>
</tr>
<tr>
<td>BLAT</td>
<td>BLAST-like alignment tool</td>
</tr>
<tr>
<td>BRAGGSS</td>
<td>Biologics in RA Genetics and Genomics Study Syndicate</td>
</tr>
<tr>
<td>BSRBR</td>
<td>British society of rheumatologists biologics register</td>
</tr>
<tr>
<td>CCD</td>
<td>Charge coupled device</td>
</tr>
<tr>
<td>CGI</td>
<td>CpG island</td>
</tr>
<tr>
<td>ChIP</td>
<td>Chromatin immunoprecipitation</td>
</tr>
<tr>
<td>CLYBL</td>
<td>Citrate lyase beta like</td>
</tr>
<tr>
<td>CpG</td>
<td>C-G dinucleotides</td>
</tr>
<tr>
<td>CRP</td>
<td>C-reactive protein</td>
</tr>
<tr>
<td>CYP</td>
<td>Cytochrome P450</td>
</tr>
<tr>
<td>DAPK1</td>
<td>Death-associated protein kinase 1</td>
</tr>
<tr>
<td>DAS28</td>
<td>28-point Disease Activity Score</td>
</tr>
<tr>
<td>ddNTPs</td>
<td>Dideoxynucleotide triphosphate</td>
</tr>
<tr>
<td>DMARD</td>
<td>Disease modifying anti rheumatic drug</td>
</tr>
<tr>
<td>DMH</td>
<td>Differential methylation hybridisation</td>
</tr>
<tr>
<td>DMP</td>
<td>Differentially methylated position</td>
</tr>
<tr>
<td>DMR</td>
<td>Differentially methylated region</td>
</tr>
<tr>
<td>DMSO</td>
<td>Dimethyl sulfoxide</td>
</tr>
<tr>
<td>Abbreviation</td>
<td>Full Form</td>
</tr>
<tr>
<td>--------------</td>
<td>-----------</td>
</tr>
<tr>
<td>DNMTs</td>
<td>DNA methyltransferase enzymes</td>
</tr>
<tr>
<td>dNTP</td>
<td>Deoxyribonucleotide triphosphate</td>
</tr>
<tr>
<td>DZ</td>
<td>Dizygotic</td>
</tr>
<tr>
<td>EBV</td>
<td>Epstein Barr virus</td>
</tr>
<tr>
<td>EDTA</td>
<td>Ethylenediaminetetraacetic acid</td>
</tr>
<tr>
<td>ENCODE</td>
<td>Encyclopedia of DNA Elements</td>
</tr>
<tr>
<td>ESR</td>
<td>Erythrocyte sedimentary rate</td>
</tr>
<tr>
<td>EULAR</td>
<td>European league against rheumatism</td>
</tr>
<tr>
<td>EWAS</td>
<td>Epigenome-wide association study</td>
</tr>
<tr>
<td>EYA4</td>
<td>Eyes absent homolog 4</td>
</tr>
<tr>
<td>EZH2</td>
<td>Enhancer of zeste homologue 2</td>
</tr>
<tr>
<td>FACS</td>
<td>Fluorescence activated cell sorting</td>
</tr>
<tr>
<td>Fc</td>
<td>Fragment crystallisable</td>
</tr>
<tr>
<td>FDR</td>
<td>False discovery rate</td>
</tr>
<tr>
<td>FLS</td>
<td>Fibroblast-like synoviocytes</td>
</tr>
<tr>
<td>FMS</td>
<td>Fragmentation solution</td>
</tr>
<tr>
<td>GWAS</td>
<td>Genome-wide association study</td>
</tr>
<tr>
<td>H3K27ac</td>
<td>Acetylation of histone H3 at lysine 27</td>
</tr>
<tr>
<td>H3K4me1</td>
<td>Monomethylation of histone H3 at lysine 4</td>
</tr>
<tr>
<td>H3K4me3</td>
<td>Trimethylation of histone H3 at lysine 4</td>
</tr>
<tr>
<td>HACA</td>
<td>Human anti-chimeric antibodies</td>
</tr>
<tr>
<td>HAQ</td>
<td>Health assessment questionnaire</td>
</tr>
<tr>
<td>HAT</td>
<td>Histone deacetylase transferases</td>
</tr>
<tr>
<td>HDAC</td>
<td>Histone deacetylase</td>
</tr>
<tr>
<td>HEPES</td>
<td>4-(2-hydroxyethyl-1-piperidineethanesulfonic acid</td>
</tr>
<tr>
<td>HIV</td>
<td>Human immunodeficiency virus</td>
</tr>
<tr>
<td>HLA</td>
<td>Human leukocyte antigen</td>
</tr>
<tr>
<td>HSO₃⁻</td>
<td>Sodium bisulphite</td>
</tr>
<tr>
<td>Abbreviation</td>
<td>Full Form</td>
</tr>
<tr>
<td>--------------</td>
<td>-----------</td>
</tr>
<tr>
<td>idat</td>
<td>Image data</td>
</tr>
<tr>
<td>IgG1</td>
<td>Immunoglobulin G</td>
</tr>
<tr>
<td>IL-1</td>
<td>Interleukin-1</td>
</tr>
<tr>
<td>IL-17</td>
<td>Interleukin-17</td>
</tr>
<tr>
<td>IL-6</td>
<td>Interleukin-6</td>
</tr>
<tr>
<td>JAK</td>
<td>Janus Kinase</td>
</tr>
<tr>
<td>JIA</td>
<td>Juvenile idiopathic arthritis</td>
</tr>
<tr>
<td>LiMS</td>
<td>Laboratory information management system</td>
</tr>
<tr>
<td>LRP1</td>
<td>Low density lipoprotein associated protein 1</td>
</tr>
<tr>
<td>LRPAP1</td>
<td>Low density lipoprotein receptor-related protein associated protein 1</td>
</tr>
<tr>
<td>LRRC17</td>
<td>Leucine rich repeat containing 17</td>
</tr>
<tr>
<td>MA1</td>
<td>Multi-sample amplification 1 mix</td>
</tr>
<tr>
<td>MACS</td>
<td>Magnetic-activated cell sorting</td>
</tr>
<tr>
<td>MATURA</td>
<td>Maximising therapeutic utility for rheumatoid arthritis</td>
</tr>
<tr>
<td>MBD</td>
<td>Methyl-CpG-binding domain</td>
</tr>
<tr>
<td>MCA</td>
<td>Methylated CpG island amplification</td>
</tr>
<tr>
<td>MDR1</td>
<td>Multidrug resistance 1</td>
</tr>
<tr>
<td>MeCP2</td>
<td>Methyl-CpG-binding protein 2</td>
</tr>
<tr>
<td>meQTLs</td>
<td>Methylation quantitative trait loci</td>
</tr>
<tr>
<td>MGMT</td>
<td>06-methyl guanine-DNA methyl transferase</td>
</tr>
<tr>
<td>MHC</td>
<td>Major histocompatibility complex</td>
</tr>
<tr>
<td>miRNA</td>
<td>Micro RNA</td>
</tr>
<tr>
<td>MMP-1</td>
<td>Matrix metalloproteinase-1</td>
</tr>
<tr>
<td>MRP8/14</td>
<td>Myeloid related protein 8/14</td>
</tr>
<tr>
<td>MVPs</td>
<td>Methylation variable positions</td>
</tr>
<tr>
<td>MZ</td>
<td>Monozygotic</td>
</tr>
<tr>
<td>NADPH</td>
<td>Nicotinamide adenine dinucleotide phosphate</td>
</tr>
<tr>
<td>Abbreviation</td>
<td>Definition</td>
</tr>
<tr>
<td>--------------</td>
<td>------------</td>
</tr>
<tr>
<td>NaOH</td>
<td>Sodium hydroxide</td>
</tr>
<tr>
<td>NCF2</td>
<td>NADPH oxidase cytosolic component p67&lt;sub&gt;phox&lt;/sub&gt;</td>
</tr>
<tr>
<td>NFκB</td>
<td>Nuclear factor kappa-light-chain-enhancer of activated B cells</td>
</tr>
<tr>
<td>NGS</td>
<td>Next generation sequencing</td>
</tr>
<tr>
<td>NICE</td>
<td>National institute for health and clinical excellence</td>
</tr>
<tr>
<td>NIH</td>
<td>National Institutes of Health</td>
</tr>
<tr>
<td>NOAR</td>
<td>Norfolk Arthritis Register</td>
</tr>
<tr>
<td>NRHV</td>
<td>National Repository of Healthy Volunteers</td>
</tr>
<tr>
<td>PADI4</td>
<td>Protein-arginine deiminase type 4</td>
</tr>
<tr>
<td>PB1</td>
<td>Reagent used to prepare BeadChips for hybridisation</td>
</tr>
<tr>
<td>PB2</td>
<td>Humidifying buffer used during hybridisation</td>
</tr>
<tr>
<td>PBMC</td>
<td>Peripheral blood mononuclear cells</td>
</tr>
<tr>
<td>PCA</td>
<td>Principal components analysis</td>
</tr>
<tr>
<td>PCR</td>
<td>Polymerase chain reaction</td>
</tr>
<tr>
<td>PDZD2</td>
<td>PDZ domain containing protein 2</td>
</tr>
<tr>
<td>PDZD8</td>
<td>PDZ containing protein 8</td>
</tr>
<tr>
<td>PLAGL2</td>
<td>Pleomorphic adenoma gene-like 2</td>
</tr>
<tr>
<td>PLG</td>
<td>Phase-lock gel</td>
</tr>
<tr>
<td>PM1</td>
<td>Precipitation solution</td>
</tr>
<tr>
<td>POFUT1</td>
<td>Protein O-fucosyltransferase 1</td>
</tr>
<tr>
<td>PPI</td>
<td>Pyrophosphate</td>
</tr>
<tr>
<td>PREVeNT RA</td>
<td>Pre-clinical evaluation of novel targets in RA</td>
</tr>
<tr>
<td>PRKCZ</td>
<td>Protein kinase C, zeta</td>
</tr>
<tr>
<td>PsA</td>
<td>Psoriatic arthritis</td>
</tr>
<tr>
<td>PTPN22</td>
<td>Protein tyrosine phosphatise non-receptor type 22 gene</td>
</tr>
<tr>
<td>PTPRC</td>
<td>Protein tyrosine phosphatase receptor type C</td>
</tr>
<tr>
<td>RA</td>
<td>Rheumatoid arthritis</td>
</tr>
<tr>
<td>RA1</td>
<td>Resuspension, hybridisation and wash solution</td>
</tr>
<tr>
<td>Abbreviation</td>
<td>Description</td>
</tr>
<tr>
<td>--------------</td>
<td>-------------</td>
</tr>
<tr>
<td>RATS</td>
<td>Rheumatoid arthritis twins study</td>
</tr>
<tr>
<td>RF</td>
<td>Rheumatoid factor</td>
</tr>
<tr>
<td>RLGS</td>
<td>Restriction landmark genome scanning</td>
</tr>
<tr>
<td>RPM</td>
<td>Revolutions per minute</td>
</tr>
<tr>
<td>SDS</td>
<td>Sodium dodecyl sulfate</td>
</tr>
<tr>
<td>SE</td>
<td>Shared epitope</td>
</tr>
<tr>
<td>SENP1</td>
<td>Sentrin-specific protease 1</td>
</tr>
<tr>
<td>SF</td>
<td>Synovial fibroblast</td>
</tr>
<tr>
<td>SLE</td>
<td>Systemic lupus erythematosus</td>
</tr>
<tr>
<td>SNP</td>
<td>Single nucleotide polymorphism</td>
</tr>
<tr>
<td>SS</td>
<td>Sjogren's syndrome</td>
</tr>
<tr>
<td>STAT4</td>
<td>Signal transducer and activator of transcription 4</td>
</tr>
<tr>
<td>STM</td>
<td>Superior two-colour master mix</td>
</tr>
<tr>
<td>SWAN</td>
<td>Subset-quantile Within Array Normalisation</td>
</tr>
<tr>
<td>T1D</td>
<td>Type 1 diabetes</td>
</tr>
<tr>
<td>T2D</td>
<td>Type 2 diabetes</td>
</tr>
<tr>
<td>TEM</td>
<td>Two-colour extension master mix</td>
</tr>
<tr>
<td>TLRs</td>
<td>Toll-like receptors</td>
</tr>
<tr>
<td>TNF</td>
<td>Tumor necrosis factor</td>
</tr>
<tr>
<td>TNFR</td>
<td>Tumor necrosis factor receptor</td>
</tr>
<tr>
<td>TNFR-1</td>
<td>Tumor necrosis factor receptor 1</td>
</tr>
<tr>
<td>UTR</td>
<td>Untranslated region</td>
</tr>
<tr>
<td>WB</td>
<td>Whole blood</td>
</tr>
<tr>
<td>XC1</td>
<td>Xstain BeadChip solution 1</td>
</tr>
<tr>
<td>XC2</td>
<td>Xstain BeadChip solution 2</td>
</tr>
<tr>
<td>XC3</td>
<td>Xstain BeadChip solution 3</td>
</tr>
<tr>
<td>XC4</td>
<td>Xstain BeadChip solution 4</td>
</tr>
<tr>
<td>ZNF74</td>
<td>Zinc finger protein 74</td>
</tr>
</tbody>
</table>
The Author

Amy Webster BSc

I graduated from The University of Manchester with an honours degree in Genetics in 2011. I developed a strong interest in epigenetics during my degree, and was delighted to find a PhD project that would allow me to explore the area in more depth. During my PhD I have developed a wider understanding of the field of epigenetics, alongside the practical and analytical skills necessary to investigate epigenetic phenomena in disease. As well as the projects I have completed investigating the role of epigenetics in rheumatoid arthritis susceptibility and treatment response, during my PhD I have also relished the opportunity to meet other epigeneticists at meetings and conferences, and to learn about other areas of the field which I have not yet had the chance to research. Through this interest I have also established the ‘Northern Epigenomics Club’, a group which meets biannually to discuss their epigenetic research. The meetings have attracted a growing following, and several esteemed epigenetic researchers have come to present their work, alongside more junior researchers. Following my PhD, I plan to go travelling around India and Nepal, after which I have been offered a research fellow position at Massachusetts General Hospital and Harvard Medical School, investigating epigenetic engineering techniques. I am looking forward to exploring the field of epigenetics further, and believe the knowledge and research skills I have gained during this PhD have been instrumental in launching my career in epigenetic research.
Publications in preparation arising from this work

A. Webster, F. Zufferey, D Plant, A. Barton, J. Bell, F. Williams, J. Worthington: DNA methylation associated with rheumatoid arthritis in disease discordant monozygotic twins

A. Webster, J. Worthington, A. Barton, D. Plant: Cryopreservation of cells does not substantially alter the DNA methylome of CD3+CD4+ T cells

D. Plant*, A. Webster*, S. Eyre, M. Lunt, K.L.Hyrich, A.G. Wilson, A. W. Morgan, J. Isaacs, J. Worthington, A. Barton: Differential methylation as a biomarker of response to etanercept in patients with rheumatoid arthritis *these authors contributed equally to the work

Conference proceedings (oral presentations)

American Society of Rheumatology (Boston - November 2014)

A. Webster, F. Zufferey, D Plant, A. Barton, J. Bell, F. Williams, J. Worthington: Differential DNA methylation associated with rheumatoid arthritis in disease discordant monozygotic twins

EULAR conference (Paris- June 2014)

A. Webster: Potential anti-TNF response biomarker identified (press conference presentation)

A. Webster, D. Plant, S. Eyre, G. Wilson, A. Morgan, J. Isaacs, J. Worthington, A. Barton: Differential DNA methylation related to response to adalimumab and etanercept in patients with rheumatoid arthritis

EULAR conference (Madrid- June 2013)


2nd Annual Infinium HumanMethylation450 Array Workshop (London- April 2013)

A. Webster: Differential methylation related to response to etanercept in patients with rheumatoid arthritis
Acknowledgements

I would like to thank my supervisors, Anne Barton, Jane Worthington and Darren Plant for their continued support and encouragement, and for being better mentors than I ever could have hoped for. For the last three years they have helped me become the scientist I hoped to be, and encouraged me to achieve things I did not think possible. I would also like to thank my advisor, Steve Eyre, for his continued enthusiasm for epigenetics, and for welcoming impromptu brainstorming sessions.

I have really enjoyed my PhD, and a large contributing factor to that was the brilliant working environment of the Arthritis Research UK Centre for Genetics and Genomics. I have made lifelong friends, and learnt about a great deal more than science during my time here. I would like to thank all the staff who make this such a great place to work, particularly the PhD students, who have provided three years of calming words, support and laughter.

I would like to thank my family, particularly my parents, for always providing encouraging words and having unwavering confidence in my abilities. Finally, I would like to thank my partner, Ryan Davies, who has supported me in every endeavour and had more confidence in me than I had in myself. He has been unshakably calm and encouraging, and has never failed to provide perspective.

This work was funded by BeTheCure and supported by Arthritis Research UK.
1. Introduction
1.1. Rheumatoid arthritis

Rheumatoid arthritis (RA) is a severely disabling autoimmune disease, which, if untreated, results in irreversible joint damage (McInnes and Schett 2011). RA is characterised by synovial joint inflammation, which can cause damage to cartilage, bones and tendons. Systemic inflammation is not uncommon in RA and can result in extra articular complications, such as lung fibrosis, rheumatoid nodules and eye involvement (scleritis and episcleritis) (McInnes and Schett 2011). A characteristic of RA is the presence of autoantibodies (specifically, rheumatoid factor (RF) and anti-citrullinated peptide antibodies (ACPA)) which are used in the classification of the disorder (McInnes and Schett 2011). RA has both social and economic consequences due to the disabling nature of the disease. People with RA have a reduced quality of life and their work life is often affected, which, along with expensive treatment costs, represents a huge economic burden (Firestein 2003). Consequently, RA is an important disease to research, especially in the context of improving treatments and optimising response to therapy.

1.1.1. Epidemiology

Several studies across Europe have estimated the prevalence of RA to be between 0.5 and 1% (Aho et al. 1998; Carmona et al. 2002; Cimmino et al. 1998; Kvien et al. 1997; Power et al. 1999; Riise, Jacobsen, and Gran 2000; Saraux et al. 1999; Simonsson et al. 1999), though this has been shown to differ quite significantly in non-European populations (Silman and Pearson 2002). Prevalence of RA is approximately three times higher in women, and increases with age (Figure 1.1) (Buckwalter and Lappin 2000; Uhlig et al. 1998). Studies have shown annual incidence rates of 36 in 100,000 females and 14 in 100,000 males (Soderlin et al. 2002; Symmons et al. 1994; Uhlig et al. 1998). The general incidence rate of RA has been shown by several studies to be decreasing, an effect which is most
apparent in females (Doran et al. 2002; Enzer et al. 2002; Jacobsson et al. 1994; Kaipiainen-Seppanen et al. 1996; Shichikawa et al. 1999; Symmons et al. 2002; Uhlig and Kvien 2005). This may be due to the increased use of the oral contraceptive pill which has been demonstrated to have a protective effect (Symmons et al. 1997). This explanation is supported by the age distribution of people with RA, as the decreased rate in women is not seen in the age bracket who would be too old to have taken the oral contraceptive pill (Symmons and Harrison 2000).

Figure 1.1: Incidence of rheumatoid arthritis according to age and gender (Uhlig et al. 1998).
1.1.2. Clinical features

Due to the complexity of RA, it can be difficult to differentiate it from other forms of inflammatory arthritis, particularly in the early stages, and no single, diagnostic test for RA exists. Clinical indicators of RA include morning stiffness and symmetrical swelling of joints, however these are also indicators of other disorders and are not sufficient for conclusive classification of RA (Grassi et al. 1998). Detection of autoantibodies such as RF and ACPA are often used to aid in classification of RA, though this is not always reliable as RF is not specific to RA while ACPA, though more specific, is not always present in patients with RA (Pruijn, Wiik, and van Venrooij 2010). Classification criteria were established in 1987 to aid in the identification of RA for clinical studies, though these criteria were criticised for being unsuitable for classification of early disease due to the lack of sensitivity of the measures (Aletaha et al. 2010). These guidelines were refined in 2010 resulting in seven clinical criteria; swelling (synovitis) in at least one joint, with the synovitis unexplained by another disorder; joint involvement including small joints; positive serology for RF and/or ACPA; an abnormal erythrocyte sedimentary rate (ESR) or abnormal c-reactive protein (CRP), and symptoms persisting for more than 6 weeks (Aletaha et al. 2010). The updated guidelines are summarised below in Table 1.1. For a definite classification as RA, patients must have a score of 6 or more using the scoring system described in Table 1.1.
## Classification Criteria for RA

<table>
<thead>
<tr>
<th>Joint Involvement</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 large joint</td>
<td>0</td>
</tr>
<tr>
<td>2-10 large joints</td>
<td>1</td>
</tr>
<tr>
<td>1-3 small joints (with or without involvement of large joints)</td>
<td>2</td>
</tr>
<tr>
<td>4-10 small joints (with or without involvement of large joints)</td>
<td>3</td>
</tr>
<tr>
<td>More than 10 joints (at least one small joint)</td>
<td>5</td>
</tr>
</tbody>
</table>

### Serology

<table>
<thead>
<tr>
<th></th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Negative RF and negative ACPA</td>
<td>0</td>
</tr>
<tr>
<td>Low-positive RF or low-positive ACPA</td>
<td>2</td>
</tr>
<tr>
<td>High-positive RF or high-positive ACPA</td>
<td>3</td>
</tr>
</tbody>
</table>

### Acute-phase reactants (at least one test results needed for classification)

<table>
<thead>
<tr>
<th></th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Normal CRP and normal ESR</td>
<td>0</td>
</tr>
<tr>
<td>Abnormal CRP or normal ESR1</td>
<td>1</td>
</tr>
</tbody>
</table>

### Duration of Symptoms

<table>
<thead>
<tr>
<th></th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Less than 6 weeks</td>
<td>0</td>
</tr>
<tr>
<td>More than 6 weeks</td>
<td>1</td>
</tr>
</tbody>
</table>

Table 1.1: Summary of the 2010 classification criteria for RA from the American College of Rheumatology/European League Against Rheumatism (EULAR). This criteria applies to newly presenting patients who have clinical synovitis in at least one joint which cannot be explained by another disease (Aletaha et al. 2010)
1.1.3. **Pathogenesis**

RA is an autoimmune disorder, meaning that during disease progression, the immune system loses tolerance for ‘self’ tissue (McInnes and Schett 2011). This loss of tolerance causes inflammatory responses leading to synovitis of the joints (Figure 1.2). This immune response leads to destruction of cartilage and bone in the joints, which if left untreated eventually causes the characteristic deformity of joints in RA (McGonagle et al. 2009; Visser et al. 2002). Production of autoantibodies such as RF and ACPA are characteristic of RA. RF is usually an IgG or IgM antibody against the subject’s own IgG molecule and may be induced during infection by the formation of immune complexes (Sattar and McInnes 2005). The reason the immune system loses tolerance for the bodies’ own tissue is not fully understood, though it may be due to post-transcriptional modifications of proteins, such as citrullination, which may prevent the immune system from recognising tissue from the organisms own body (De et al. 2004; Vincent et al. 1999). Both RF and ACPA can be present for many years before onset of the disease, with ACPA appearing earlier, although whether the autoantibodies are directly involved in pathogenesis remains an active area of research (reviewed in (Burska et al. 2014)).

Adaptive immune pathways drive early RA pathogenesis. These involve the production of cytokines including tumor necrosis factor α (TNF-α), interleukin-1 (IL-1) and interleukin-6 (IL-6), which activate cells important in RA pathogenesis including leukocytes, endothelial cells, synovial fibroblasts and osteoclasts (Brennan and McInnes 2008; Feldmann, Brennan, and Maini 1996; Hess et al. 2011). TNF-α has both local effects in RA pathogenesis, including activation of monocytes and increasing cytokine production in endothelial cells (Feldmann, Brennan, and Maini 1996; Hess et al. 2011), and systemic consequences such as dysregulation of the hypothalamic pituitary adrenal axis and increased cardiovascular disease risk (Sattar et al. 2003). IL-6 activates osteoclasts and promotes proliferation of T-cells
and B-cells. It is also involved in pannus formation (Nakahara et al. 2003; Paleolog 2002) and has widespread systemic effects including osteoporosis and anaemia (De et al. 2006; Nemeth et al. 2004). IL-1 increases synovial and monocyte cytokine production and activates osteoclasts (Brennan and McInnes 2008; McInnes and Schett 2007) alongside systemic effects similar to TNF-α.

Many cells are known to influence development of RA, particularly T cells and B lymphocytes. B cells are involved in pathogenesis of RA through production of cytokines and autoantibodies. Synovitis is known to be promoted by interleukin-17 (IL-17), a cytokine produced during the differentiation of Th17 cells (Nalbandian, Crispin, and Tsokos 2009) which is involved in recruitment of monocytes and neutrophils, and facilitates activation and infiltration of T-cells (Nalbandian, Crispin, and Tsokos 2009).

Figure 1.2: Rheumatoid arthritis affected joint compared to healthy joint, showing inflamed synovium.
1.1.4. Comorbidity

RA is also associated with systemic complications such as cardiovascular, psychological and skeletal disorders (McInnes and Schett 2011). The main systemic consequence of RA is an increase in cardiovascular illness including myocardial infarction, cerebrovascular events and heart failure (Holmqvist et al. 2010; Solomon et al. 2003; Wolfe et al. 1994). Other systemic consequences in major organs which are affected by RA include reduced cognitive function in the brain, liver complications, inflammatory and fibrotic disease in the lungs and sarcopenia of the muscles (McInnes and Schett 2011).

1.1.5. Economic cost of rheumatoid arthritis

Due to the debilitating nature and the widespread occurrence of RA, it represents a huge economic burden on healthcare systems and the general economy. Estimates from the National Audit Office showed the cost to the health system in England incurred due to RA is approximately £560 million annually, while the cost of work related disability was estimated to be an additional £1,800 million annually (National Audit Office 2009). These figures show the importance of developing effective treatments for RA and justify the use of expensive drugs. Indeed, it has been consistently shown that early, effective treatment of RA can prevent future disability and improve long-term outcomes, including work disability (Demoruelle and Deane 2012; Emery and Salmon 1995; Gough et al. 1994).
1.2. **Treatment of rheumatoid arthritis**

RA is traditionally managed by treating symptoms using corticosteroids and anti-inflammatory drugs (Gaffo, Saag, and Curtis 2006). These are usually combined with disease modifying anti-rheumatic drugs (DMARDs) to slow disease progression (Gaffo, Saag, and Curtis 2006).

The effectiveness of treatment in RA is measured using a combination of clinical measurements such as swollen and tender joint counts, ESR (or CRP) and measures of quality of life of the patient (Scott 2012). These variables are usually combined into a clinical index of disease activity, such as the 28 joint count Disease Activity Score (DAS28) (Fransen and van Riel 2005). According to the EULAR response criteria, the change in DAS28 shown over the course of a treatment represents the level of an individuals’ response to a treatment (Fransen and van Riel 2005).

1.2.1. **DMARDs**

DMARDs are generally very effective in treating RA, as they reduce synovitis, systemic inflammation and overall disability (Donahue et al. 2008). The most commonly used DMARD is methotrexate, though other DMARDs such as sulfasalazine and leflunomide are also regularly prescribed (Scott 2012). They are often used in combination, which has a greater clinical effect, however approximately a third of patients do not respond to treatment with DMARDs (Nagashima *et al*. 2006; Scott 2012). Some patients suffer minor adverse events such as nausea, and rare but serious adverse events such as blood and liver toxicity are also associated with use of DMARDs (Scott 2012).
1.2.2. **Biologics**

Biologics are a relatively new form of treatment for RA and represent a huge change in the way in which RA is treated. Biologics are genetically engineered human proteins which target specific large molecules involved in inflammatory pathogenesis, such as tumour necrosis factor (TNF), IL-1 and IL-6 (Charles *et al.* 1999; Feldmann, Brennan, and Maini 1996). The lack of specificity in previous drug classes can have serious side effects, which biologics, by targeting specific components, should avoid (Scott 2012).

There are different categories of biologic drugs, which target and interact with different components of the immune system in order to reduce the inflammatory response, as shown in Figure 1.3. The most widely used biologic drugs are TNF antagonists such as etanercept, infliximab, adalimumab, certolizumab pegol and golimumab (Wilkie and Schwieterman 2012). Other types of biologic drugs include IL-1 inhibitors such as anakinra and IL-6 inhibitors such as tocilizumab (Wilkie and Schwieterman 2012). There are also biologic drugs which interact with other components of the immune system, including abatacept, which blocks T-cell costimulation, and rituximab, which depletes B-cells (Wilkie and Schwieterman 2012).

In order for patients to be prescribed biologic drug therapies in the UK, they must fulfil the National Institute for Health and Care Excellence (NICE) criteria: they must have been classified as having RA according to the American College of Rheumatology (ACR) 1987 criteria; they must have active RA with a DAS28 score of more than 5.1, and they must have failed therapy with methotrexate and at least one other standard DMARD (Ledingham and Deighton 2005).
Figure 1.3: Interaction between different classes of biologic drugs and the inflammatory pathway in rheumatoid arthritis.
1.2.3. **Anti-TNF therapies**

TNF is a cytokine which is important in host defence and control of tumour growth (Wilkie and Schwieterman 2012). TNF causes inflammation by inducing proinflammatory cytokines, including IL-1 and IL-6, which are involved in RA pathogenesis, and increasing prostaglandin synthesis (Wilkie and Schwieterman 2012). There are several members of the TNF family, the most important of which are TNF and lymphotoxin (previously known as TNF-β). TNF is biologically active in both membrane-bound and soluble forms (Wilkie and Schwieterman 2012).

TNF was revealed as a possible therapeutic target due to its high levels of expression within the synovial membrane, along with its two receptors, p55 and p75 TNFR (Chu et al. 1991; Husby and Williams, Jr. 1988). Subsequently it was found that by neutralising TNF in a cell culture of RA synovial cells, the production of IL-1, IL-6, IL-8 and GM-CSF were all reduced (Brennan et al. 1989; Butler et al. 1995; Haworth et al. 1991). Later observations of the pathogenic effect of TNF in transgenic mice and the therapeutic effects of monoclonal anti-human-TNF showed TNF was potentially a very important therapeutic target for RA (Brennan, Maini, and Feldmann 1992; Feldmann et al. 1990; Keffer et al. 1991).

There are currently five biologic drugs available that act as TNF inhibitors, whose properties are summarised below in Table 1.2 (Scott 2012). Comparison of the efficacies of these drugs through a mixed treatment comparison found etanercept to be the most efficacious (Schmitz et al. 2012). All the anti-TNF biologic drugs except infliximab are administered by subcutaneous injection, the frequency of which is determined by the drug’s half-life (Wilkie and Schwieterman 2012).

Due to the role of TNF in the immune response, its inhibition has quite serious but rare side effects typically associated with immune suppression including increased risk of infections such as listeria, histoplasmosis and aspergillosis (the main adverse
effects of anti-TNF therapies are summarised in Figure 1.4) (Warris, Bjorneklett, and Gaustad 2001). TNF also plays a pivotal role in control of tumour growth, and consequently its inhibition was previously associated with a dose-dependent increase in risk of malignancies (Bongartz et al. 2006); however a more recent study did not identify an increased risk of solid organ cancer or lymphoma in patients treated with anti-TNF therapy (Mercer et al. 2014).

<table>
<thead>
<tr>
<th>Chemical Name</th>
<th>Commercial Name</th>
<th>Mechanism of TNF Inhibition</th>
<th>Half Life</th>
</tr>
</thead>
<tbody>
<tr>
<td>Etanercept</td>
<td>Enbrel</td>
<td>Circulating receptor fusion protein</td>
<td>4 days</td>
</tr>
<tr>
<td>Infliximab</td>
<td>Remicade</td>
<td>Chimeric monoclonal antibody</td>
<td>~9 hours</td>
</tr>
<tr>
<td>Adalimumab</td>
<td>Humira</td>
<td>Recombinant monoclonal antibody</td>
<td>14 days</td>
</tr>
<tr>
<td>Certolizumab</td>
<td>Cimzia</td>
<td>Recombinant monoclonal antibody</td>
<td>14 days</td>
</tr>
<tr>
<td>Golimumab</td>
<td>Simponi</td>
<td>Recombinant monoclonal antibody</td>
<td>14 days</td>
</tr>
</tbody>
</table>

Table 1.2: TNF Inhibitors used in the treatment of rheumatoid arthritis (adapted from (Scott 2012; Wilkie and Schwieterman 2012)).
Figure 1.4: Main adverse effects associated with use of anti-TNF therapy (Willrich, Murray, and Snyder 2014).
1.2.3.1. **Etanercept**

Etanercept is a TNF-receptor fusion protein which interacts with TNF in its soluble form. It is composed of two dimers, each containing a TNF-receptor protein and a portion of human immunoglobulin 1 (IgG1) (Scott 2012; Wilkie and Schwieterman 2012). It binds both TNF and lymphotoxin, preventing interaction with their receptors (Scott 2012). Since its half life is 4 days, etanercept is administered either once (50mg) or twice (25mg) weekly (Scott 2012). Since etanercept is administered by injection, a common side effect experienced by patients is a reaction at the site of injection (Khraishi 2009).

1.2.3.2. **Infliximab**

Infliximab is a chimeric IgG1 antibody which binds TNF in both membrane bound and soluble forms (Scott 2012). It is a chimeric antibody in that the antigen binding region is derived from a mouse while the constant region is derived from a human antibody (Scott 2012). Infliximab prevents TNF from binding to its receptor and also has a cytotoxic effect on cells which express TNF (Scott 2012). Unlike the other biologic drug treatments used in RA, infliximab is administered intravenously. The typical dose is 3mg/kg every 8 weeks; however there is a great range of responses between patients, so many require either larger or more frequent doses (St Clair et al. 2002). Because of the mode of administration, in addition to the expected side effects of biologics, some people suffer infusion reactions such as headache and nausea (Wilkie and Schwieterman 2012). Secondary non-response has been noted in approximately half of patients treated with infliximab within their first year of treatment, which may be due to the development of anti-treatment antibodies, such as human anti-chimeric antibodies (HACA) (Buch et al. 2007).
1.2.3.3. **Adalimumab**

Adalimumab is a fully humanised recombinant IgG1 monoclonal antibody, and is consequently less immunogenic than infliximab (Wilkie and Schwieterman 2012). Adalimumab inhibits TNF from binding to its receptors and lyases cells expressing surface TNF (Scott 2012). As well as improving symptoms of RA, adalimumab has been shown to slow disease progression (Wilkie and Schwieterman 2012). It has a rapid clinical effect, though there are significant differences in response rate among patients (Scott 2012; Wilkie and Schwieterman 2012). When combined with the DMARD methotrexate, adalimumab is much more effective (Breedveld et al. 2006).

1.2.3.4. **Certilizumab**

Certilizumab pegol is a recombinant monoclonal antibody. It is made up of a humanised antigen-binding domain of the TNF antibody joined to a polyethylene glycol molecule. This extends the plasma life of the molecule to approximately 2 weeks. It targets both membrane-bound and soluble TNF (Scott 2012; Wilkie and Schwieterman 2012). Unlike the other anti-TNF biologics, certolizumab does not have a fragment crystallisable (Fc) region and thus does not cause cell mediated cytotoxicity (Nesbitt et al. 2007; Scott 2012).

1.2.3.5. **Golimumab**

Golimumab is a human derived IgG1 monoclonal antibody produced in a transgenic mouse. It binds to TNF in both soluble and membrane bound form. It is administered monthly, initially at 50mg which can be increased after 4 doses to 100mg if there is no response (Scott 2012; Wilkie and Schwieterman 2012).
1.2.4. **Future treatment of rheumatoid arthritis**

Emerging biologic therapies are being developed to target other pathways of RA pathogenesis, with the aim of achieving improved efficacy compared to existing biologics. Monoclonal antibodies are being developed to target proinflammatory cytokines, including canakinumab and gevokizumab which target IL-1 (Alten et al. 2011; Geiler and McDermott 2010), sirukumab which targets IL-6 (Xu et al. 2011) and secukinumab which targets IL-17 (Genovese et al. 2013; Hueber et al. 2010). B-cell depleting therapies including ofatumumab (Ostergaard et al. 2010; Taylor et al. 2011) and atacicept (Genovese et al. 2011b; Tak et al. 2008; van Vollenhoven et al. 2011) are also being developed.

Therapies targeting small molecules are also being developed in the hope that their oral administration will make them cheaper than current biologic therapies. Therapies are currently being developed to inhibit intracellular signalling molecules, most notably tofacitinib which inhibits Janus kinase, an intracellular tyrosine kinase (Kremer et al. 2012; Tanaka et al. 2011). Fostamatinib is another tyrosine kinase inhibitor being developed which targets spleen tyrosine kinase (Genovese et al. 2011a; Weinblatt et al. 2008; Weinblatt et al. 2010).
1.2.5. Key research questions

RA is a common complex autoimmune disorder with serious long-term sequelae if left untreated. A range of effective therapies are available but each has a variable response rate in patients. As the introduction of effective therapy early in the disease course is the best predictor of long-term outcome, selecting the right treatment for the right patient early in disease is a priority. Thus, two of the most important research questions are, first, what are the risk factors for disease development and, second, how can the right therapy be selected for individual patients earlier in the disease course? Both RA, itself, and treatment response are likely to be multifactorial traits involving both genetic and environmental risk factors, discussed in more detail below (Aho and Heliovaara 2004).
1.2.6. Non-genetic risk factors in the development of rheumatoid arthritis

Results from twin studies support a significant role for environmental triggers in determining RA risk, evidenced by high discordance rates between monozygotic twins (section 1.2.7) (Meyerowitz, Jacox, and Hess 1968; Silman, Newman, and MacGregor 1996).

Smoking was first identified as a risk factor for RA more than 20 years ago after it was found that RA is three times more common in people who smoke than non-smokers (Vessey, Villard-Mackintosh, and Yeates 1987) and since then smoking has been established as one of the most important causative and exacerbating environmental exposures in RA (Hoovestol and Mikuls 2011; Lahiri et al. 2014).

The importance of interactions between genes and the environment in determining development of RA has recently been reported. For example, the HLA-DRB1 shared epitope (section 1.2.7.1) in combination with smoking has been observed to increase the risk of developing ACPA positive RA by 21 fold (95%CI 11-40.2) (Karlson et al. 2010; Klareskog et al. 2006).

Breastfeeding has also been implicated as a risk factor in the development of RA, as women who are exposed to breast-feeding following the birth of their first child were observed to be at higher risk of developing the disease (OR 5.4, 95%CI 2.5-11.4) compared with women who do not breastfeed (Brennan and Silman 1994). This is contradicted by a recent study that identified a dose-dependent inverse association of breastfeeding with RA and inflammatory polyarthritis (Lahiri et al. 2014). Studies from the Norfolk Arthritis Register (NOAR) found that women with inflammatory polyarthritis who had a history of live births had improved disease outcome, while those with adverse pregnancy outcomes tend to have worse disease prognosis (Camacho et al. 2011a; Camacho et al. 2012). Another study
from the NOAR cohort showed that use of oral contraceptives is associated with improved outcome (Camacho et al. 2011b).

Synovial fluid of RA patients shows evidence of viral products, indicating infection could be playing a role in disease development (Stransky et al. 1993). Various viruses have been investigated in relation to RA however the association with the greatest body of evidence is that of Epstein-Barr virus (EBV). Patients with RA have been found in several studies to have a higher level of antibodies against EBV encoded proteins when compared with healthy controls (Alspaugh et al. 1981; Blaschke et al. 2000; Catalano et al. 1979; Ferrell et al. 1981; Hazelton, Sculley, and Pope 1987). Individuals with RA have also been shown to have a higher proportion of EBV infected B cells and evidence of EBV in the synovium (Blaschke et al. 2000; Saal et al. 1999; Takeda et al. 2000; Takei et al. 1997; Tosato et al. 1984). A study from the NOAR cohort also found that a higher than expected proportion of patients presenting with symptoms had recently been exposed to tetanus immunisation (Symmons et al. 1997; Symmons and Chakravarty 1993).

Interestingly, epidemiological studies have also identified environmental factors which are protective of disease, such as alcohol consumption and high social class (Lahiri et al. 2014).

Whilst studies have identified a number of environmental factors that predispose to RA, more progress has been made in identifying genetic risk factors.
1.2.7. **Genetic risk factors for susceptibility to rheumatoid arthritis**

Twin and family studies support the link between genetics and RA. The concordance rates between monozygotic and dizygotic twins are 15% and 3-4%, respectively (Aho et al. 1986; MacGregor et al. 2000; Silman et al. 1993). Since monozygotic twins share both genes and environment, whilst dizygotic twins share environment, but only 50% of their genome, this is strong evidence for a genomic component in susceptibility to disease. Heritability estimates based on twin studies from two populations (UK and Sweden) reported figures of ~60%, which is similar to qualitative genetic analysis which has shown that the heritability of RA is approximately 60% (Aho et al. 1986; MacGregor et al. 2000; Silman et al. 1993). Geographical studies also support a genetic role in RA aetiology, as prevalence of the disease varies in different countries, though this may also indicate the role of differences in lifestyle and environmental exposures to disease onset. For example the prevalence in Pima Indians and Chippewa Indians is 5.3% and 6.8% respectively (Del et al. 1989; Harvey et al. 1981). By contrast, Chinese and Japanese populations have lower prevalence rates of 0.3% and 0.2%, respectively (Shichikawa et al. 1999; Zeng, Huang, and Chen 1997).
1.2.7.1. **HLA**

The most significant genetic association with RA is the human leukocyte antigen (HLA)-*DRB1* gene, which was identified as a risk factor for RA in 1978 after being investigated due to its role in the immune response (Stastny 1978). Since then, several alleles of *HLA-DRB1* have been identified as being associated with RA and validated in many studies in several different populations (Balsa et al. 2000; Citera et al. 2001; del, I and Escalante 1999; Pascual et al. 2001; Terao et al. 2011; Wakitani et al. 1997). The multiple *HLA-DRB1* alleles associated with RA were all found to share a short amino acid sequence in the third hypervariable region at positions 70-74 collectively termed the shared epitope (SE) (Gregersen, Silver, and Winchester 1987).

The HLA family of genes encode cell surface antigen-presenting proteins and reside on chromosome 6 within the major histocompatibility complex (MHC) (Francke and Pellegrino 1977). Recently, dense genetic mapping of the HLA region has refined the SE hypothesis (Raychaudhuri et al. 2012). Although amino acids 71 and 74 within the region are still independently associated with disease, it appears the strongest association is from amino acid 11 outside the traditional SE region, but still within the antigen binding groove. Separate associated variants encoding amino acid 9 within HLA-B, and amino acid 9 within HLA-DPβ1 are other independent risk factors for RA within the MHC region. The 5 independent risk amino acids in HLA-DRB1, HLA-B and HLA-DPβ1 account for almost all of the association signal in the HLA region (Raychaudhuri et al. 2012). The most significantly associated haplotype has an odds ratio of 4.44 (Raychaudhuri et al. 2012).
The second most important risk locus for RA is the protein tyrosine phosphatase non-receptor type 22 gene (*PTPN22*) on chromosome 1. *PTPN22* was first associated with RA in a US population, in a study examining 87 potential functional single nucleotide polymorphisms (SNPs) from within interesting candidate genes mapping under peaks of linkage identified in previous affected sibling pair family studies (Begovich et al. 2004). The rs2476601 SNP correlated with RA with an odds ratio of 1.65 ($p=6.6\text{E}-4$) in the discovery study (including 475 cases and 475 controls), and an odds ratio of 1.97 ($p=5.6\text{E}-8$) in a subsequent replication study (including 463 cases and 926 controls) (Begovich et al. 2004). This association has since been replicated in many studies, and in different populations (Farago et al. 2009; Kokkonen et al. 2007; Martin et al. 2011; PIERER et al. 2006; Plenge et al. 2005; Simkins et al. 2005; Viken et al. 2007; Wesoly et al. 2007; Zhernakova et al. 2005). The rs2476601 variant is a co-dominant allele, demonstrated by the fact that the risk for RA doubles when a single T allele is present (heterozygous) compared to individuals who are homozygous for the C allele at that locus (OR 1.5-2.0) (Bottini et al. 2004; Hinks et al. 2005; Lee et al. 2005; Simkins et al. 2005; Steer et al. 2005; van et al. 2005). The SNP is much more common in individuals who are positive for ACPA (Zheng et al. 2012) although it has been found to reach genome-wide significance in both ACPA positive and negative populations (Padyukov et al. 2011; Viatte et al. 2012). While this SNP is a strong RA susceptibility gene, once disease has developed, it does not predict disease severity (Feitsma et al. 2007; Taylor et al. 2013) or treatment response (Daien et al. 2012; Kokkonen et al. 2007; MAJORCZYK, Pawlik, and KUSNIERCZYK 2010; POTTER et al. 2009).

The gene is important in the regulation of immune cell signalling (Bottini and Peterson 2014; Rhee and Veillette 2012) including regulation of T-cell receptor
signalling thresholds. While its role in T cells is quite well defined, it also has a role in B cells and myeloid cells which is less established (Stanford and Bottini 2014).

The rs2476601 SNP causes a substitution of arginine to tryptophan in the PTPN22 protein (Bottini et al. 2004). Recent studies show conflicting evidence on the effect of the SNP. One study suggested the associated SNP is a loss of function variant, decreasing binding in the functional peptide groove, and increasing the rate at which the protein is degraded (Zhang et al. 2011; Zikherman et al. 2009). Another study, however, showed the opposite effect, suggesting the SNP causes gain of function as the phosphatase encoded by the gene has increased catalytic activity (Fiorillo et al. 2010; Rieck et al. 2007; Vang et al. 2005; Vang et al. 2012). The SNP also has consequences at a cellular level, as the number of pathogenic autoimmune T cells is increased (Vang et al. 2013).

There is substantial overlap in genetic risk in different autoimmune disorders, and PTPN22 is a good example of an overlapping gene as it is associated with type 1 diabetes (Bottini et al. 2004; Fichna et al. 2010; Korolija et al. 2009; Saccucci et al. 2008; Zheng and She 2005), systemic lupus erythematosus (Kyogoku et al. 2004; Orozco et al. 2005) and progressive systemic sclerosis (az-Gallo et al. 2011b). Interestingly, it is also associated with Crohn’s disease however it is associated with reduced risk, representing a protective effect of the minor allele (az-Gallo et al. 2011a).

While HLA-DRB1 and PTPN22 are important genetic risk factors in RA, together they account for just 40% of the total genetic risk, showing that many other genes are involved.
1.2.7.3. Other genetic risk factors for rheumatoid arthritis

Since 2007, genome wide association studies (GWAS) have emerged as a powerful tool to determine genetic association with disease by comparing genetic variants across the entire genome between affected and non affected individuals. GWAS allows the hypothesis-free testing of millions of common SNPs for association with a trait, to identify genetic markers of disease. The development of GWAS has allowed rapid progress in the identification of RA susceptibility genes.

The first major GWAS was conducted by the Wellcome Trust Case Control Consortium (Burton et al. 2007). This study investigated seven major diseases, including 2000 cases from each disease and approximately 3000 shared controls. The diseases investigated were RA, bipolar disorder, type 1 and 2 diabetes, coronary artery disease, Crohn’s disease and hypertension. The study confirmed association between HLA-DRB1 and PTPN22 and RA risk, and identified suggestive association at 9 previously unknown susceptibility loci (p-values in the range E-5 to E-7). The study also identified 49 SNPs which were putatively associated (p=E-4 to E-5). Three of these susceptibility loci were subsequently validated in an independent cohort of 4,106 RA cases and 11,238 controls (Barton et al. 2008).

Many more RA susceptibility loci have since been identified as a result of well powered GWAS and subsequent meta-analyses and by 2010, 31 susceptibility loci had been confirmed at genome-wide levels of statistical association (p<5E-8) in European populations (Plenge et al. 2007; Raychaudhuri et al. 2009; Stahl et al. 2010). In a Japanese study, further loci were identified (Okada et al. 2012).

The next major advance came in 2012 following the development of a custom SNP array that took advantage of the overlapping susceptibility loci in a number of autoimmune disorders to compile an array for dense genotyping of 186 loci, the
Immunochip array. In a study of 11,475 RA cases and 15,870 controls, 14 novel susceptibility loci were identified (Eyre et al. 2012). Five of these loci (rs2476601, rs59466457, rs6032662, rs8043085 and rs629326) were associated with RA only in ACPA-positive RA, indicating differences in genetic predisposition to antibody positive and negative disease. This disparity may also have been influenced by differences of power in the study populations, as there were around three times more ACPA positive individuals than ACPA negative, thus there was more power to detect genetic associations among the ACPA positive subset. Subsequent studies in ACPA-negative individuals, although less well-powered, identified that a number of these loci were also associated with that subgroup (Viatte et al. 2012).

Recently a well powered meta-analysis involving more than 100,000 individuals of European and Asian ancestry (29,880 RA cases) identified 42 new susceptibility loci at genome-wide significance, bringing the total number of RA susceptibility loci identified to date to 101 (Okada et al. 2014). The variance explained by these susceptibility loci is only around 17% (Figure 1.5) indicating that there is still a substantial amount of missing heritability in RA susceptibility, which is discussed in more detail in section 1.3.10.8. Interestingly, the CTLA-4, IL6R and TYK2 genes are all RA susceptibility genes and also the targets or analogues of biologic drugs used to treat RA (abatacept, tocilizumab and tofacitinib, respectively). Indeed, the susceptibility genes identified were enriched for drug targets and may indicate that therapies used in other diseases could also be appropriate for the treatment of RA. For example, CDK6 and CDK4 are RA susceptibility genes, which are targeted by existing drugs used in treatment of cancer, and could potentially be therapeutic in RA (Okada et al. 2014).
Figure 1.5: Rheumatoid arthritis genetic risk loci in order of discovery. The odds ratio and confidence interval for each locus is shown on the left Y axis and the increase in heritability explained by confirmed loci is shown on the right Y axis (Worthington, Viatte, and Eyre 2014)
1.2.7.4. Identifying causal variants from GWAS markers

While GWAS has allowed the identification of disease associated loci, the SNPs identified in these studies are markers of disease, and not necessarily the causal variant. This is due to linkage disequilibrium (LD), which is present when two or more loci are inherited together more often than would be expected by chance. Following the identification of a SNP associated with disease in GWAS, it is important to identify the variant which is causing disease. Fine mapping, meaning genotyping of multiple variants across an associated locus is usually the first step in order to refine the associated region; however it can be difficult to identify functional variants due to their LD with other variants. An example of a locus in which fine mapping was important in identifying causal variants is the SORT1 gene, which is associated with low-density lipoprotein cholesterol. Fine mapping of the gene region identified a more associated variant than that detected by the initial GWAS and consequently doubled the estimate of heritability explained by the gene (Sanna et al. 2011). For many GWAS of autoimmune diseases, the associated variants lie in intergenic, regulatory regions but it is not always obvious which gene is under the regulation of the associated variant. In order to distinguish which genes are implicated in disease pathogenesis, functional studies are necessary to identify downstream consequences of the genetic sequence changes.
1.3. Clinical predictors of response to anti-TNF therapy

1.3.1. Importance of predictors of response to anti-TNF therapy

Anti-TNF drugs represent a huge advance in the treatment of RA. Treatment of RA with anti-TNFs has been observed to suppress erosive damage to joints, providing a much better outlook and long term prognosis (Lipsky et al. 2000). However, a large proportion of patients (30-40%) have little or no response (Tan et al. 2010). The identification of patients who are likely to respond to anti-TNF drugs prior to the start of treatment would have several advantages. Anti-TNF treatment is very expensive, costing approximately $13,000 in the US and £10,000 in the UK per patient annually. By identifying and targeting for therapy those who are likely to respond, the cost effectiveness of anti-TNF treatment would be improved, making it a more viable option in the future for widespread use (Tan et al. 2010). Anti-TNF drugs are associated with potentially harmful side effects, such as increased susceptibility to infection. If patients who were likely to be non-responders to anti-TNFs could be identified prior to treatment, it would avoid unnecessary exposure to these potential side effects (Tan et al. 2010).

Ultimately, finding a marker of response which would allow the identification of potential good or poor responders prior to the start of treatment would provide a step towards stratified medicine. It would allow treatment to be allocated to patients more appropriately and allow earlier treatment with therapies which are likely to be of the most benefit to the patient, thereby providing a better long term outlook.
1.3.2. **Classifying treatment response**

Activity of disease in RA is a complex phenotype that cannot be measured by one variable, so the disease activity score (DAS) was developed to combine clinical measurements including a 44 swollen joint count, erythrocyte sedimentation rate and a measurement of general health (van der Heijde *et al.* 1990; van der Heijde *et al.* 1993). Subsequently, the DAS28 score has been developed which includes tender and swollen joint counts from 28 joints, and so is easier to measure than DAS (Prevoo *et al.* 1995). This is used in the European League Against Rheumatism (EULAR) response criteria which is used to classify people as being good, poor or moderate responders to therapy (Fransen and van Riel 2005). Good responders are defined as having a change in DAS or DAS28 of $\geq 1.2$, or having an endpoint DAS of <2.4, or an endpoint DAS28 of <3.2. Meanwhile, poor responders are defined as having a change in DAS of $< 0.6$, an endpoint DAS of $>3.7$ or an endpoint DAS28 of $>5.1$ (Fransen and van Riel 2005).

1.3.3. **Gender as an indicator of response to anti-TNF therapy**

Studies from Norway (Heiberg *et al.* 2005; Kvien *et al.* 2005; Kvien *et al.* 2006) have reported that female patients have a 30-50% less efficacious response to anti-TNF drugs compared to male patients. However, female patients who were prescribed anti-TNFs tended to have a more severe and active disease, higher DAS28 scores and delayed access to therapeutic drugs. Thus, the reduced response to anti-TNFs in female patients may be due to confounding factors which are likely to also influence drug response. This is supported by findings from the British Society of Rheumatologists Biologics Register (BSRBR), which found females were less likely to reach improvement in disease state to the point of remission (Hyrich *et al.* 2006). These findings were in contrast with those reported from a Swedish Registry which found that gender was not predictive of response to anti-TNFs in a study involving
1565 patients (Kristensen et al. 2008). Generally, males have a tendency to have a lower disease severity at baseline and appear to have higher proportion of patients achieving remission (Kristensen et al. 2008). In more severe disease, therapy is less effective thus this would give the impression of a lower drug efficacy in female patients (Kristensen et al. 2008). Because of this, it is potentially misleading to judge whether or not gender is a predicting factor in response using rate of remission (Kristensen et al. 2008) however the study from the BSRBR found that gender is associated with response independently of HAQ (Hyrich et al. 2006).

If a gender dependent difference in response to TNF inhibitors does exist, it could be explained by the differences in joint inflammation in RA between males and females. Males have higher levels of anti-inflammatory androgens in the synovial tissue than females, and some of the therapeutic effect of anti-TNFs may act through restoring androgens and blocking deleterious hormone conversions (Straub et al. 2006). Female patients also respond less well to treatment with methotrexate, showing that reduced therapeutic response in females is not specific to TNF-antagonists (Forslind et al. 2007; Kvien et al. 2006).
1.3.4. **Initial disability as an indicator of response to anti-TNF therapy**

Patients with early RA are more likely to respond to anti-TNF therapy than those who are in later stages of the disease (Emery and Dorner 2011). This is most likely due to the accumulation of irreversible structural damage in late disease (Hyrich et al. 2006). In RA, disability is often quantified using clinical measures such as health assessment questionnaires (HAQ) which measures physical disability and provides a measure of the patients perspective on the severity and impact of RA on their lives (Bruce and Fries 2005).

Low disability at baseline is generally associated with good response to anti-TNFs while high disability is associated with poor response, using the EULAR response criteria (Hyrich et al. 2006; Kristensen et al. 2008). As with patient gender, level of disability is also correlated with response to conventional DMARDs (Anderson et al. 2000).

1.3.5. **Dosage effects and adherence in anti-TNF response**

In a study performed in the Biologics in Rheumatoid Arthritis Genetics and Genomics Study Syndicate (BRAGGSS) cohort, non-adherence to subcutaneously administered anti-TNF therapies at 3 months was found to be associated with a reduced response and a worse clinical outcome at 6 months (Bluett et al. 2014). That study found that of 286 patients, 27% self-reported non-adherence to their anti-TNF, indicating the importance of research into why patients do not adhere to their therapy.
1.3.6. Concurrent use of DMARDs as an indicator of response to anti-TNF therapy

Evidence from the BSRBR and the South Swedish Arthritis Treatment Group Register shows that concurrent use of methotrexate alongside anti-TNF therapy is associated with several measures of good response rates in patients with less severe disease, including EULAR remission (DAS28<2.6) and good response, as well as ACR50 (an improvement of at least 50% from baseline) and ACR70 (an improvement of 70% from baseline) responses (Hyrich et al. 2006; Kristensen et al. 2008). This positive effect was seen at 3 and 6 months following commencement on treatment, demonstrating the long-term benefits of combined therapy (Kristensen et al. 2008). This improvement is unlikely to be due to an increased efficacy of anti-TNFs in the presence of DMARDs, but could be due to the fact that DMARDs such as methotrexate are themselves, very strong and effective drugs so are likely having a therapeutic effect which is somehow enhanced by the presence of anti-TNFs (Kristensen et al. 2008). Another possibility is that the immunosuppressive action of methotrexate renders chimeric anti-TNFs less immunogenic therefore improving outcome, however this would not explain the increased response rate of concurrent methotrexate use with etanercept as it does not produce immunogenicity (Dore et al. 2007).

Anti-TNF drugs are usually prescribed to patients who have had a poor response to methotrexate or other DMARDs in the past (Kristensen et al. 2008). This suggests there is a possibility that tolerance to methotrexate or DMARDs may be a characteristic of an underlying genetic or environmental state which also confers good treatment response to anti-TNFs (Kristensen et al. 2008).
1.3.7. **Presence of autoantibodies as an indicator of response to anti-TNF therapy**

Preliminary research has indicated that autoantibody positivity may be predictive of response to biologic therapy. Although association between autoantibody positivity and response has not been observed for all drug classes, e.g. tocilizumab (Burmester et al. 2011); several studies have reported an association with rituximab (Emery and Dorner 2011). For example, the results of controlled trials have reported that patients positive for RF and ACPA had better responses to rituximab and abatacept compared to seronegative patients (Emery et al. 2008; Pieper et al. 2013; Tak et al. 2011). However it remains unclear which autoantibody would be most useful in explaining variability in response (Isaacs and Ferraccioli 2011; Quartuccio et al. 2009). For example, one study indicated that seropositivity for either RF or ACPA (as defined by positivity for anti-CCP antibody testing) predicted a poor response to infliximab (Potter et al. 2009) whilst a separate study reported that positivity for either autoantibody predicted good response to rituximab (Strangfeld et al. 2009). A more recent study from the BRAGGSS cohort found that presence of ACPA was associated with poor response to anti-TNFs at 6 months (Fisher et al. 2012).

RF and ACPA are the autoantibodies traditionally used to classify RA, hence they were the obvious targets for prediction of drug response. However alternative autoantibodies have been suggested for this role. These include autoantibodies against other citrullinated proteins and antibodies against modified citrullinated vimentin (Bang et al. 2007; Mathsson et al. 2008).
1.3.8. **The influence of environmental factors on response to anti-TNF therapy**

Environmental factors have been found to influence drug efficacy, the most notable of which is exposure to smoking (Hyrich *et al.* 2006). Smoking has been associated with reduced response rates in patients who are being treated with infliximab (Hyrich *et al.* 2006). This might be due to a change in the pharmacokinetics of the drug caused by smoking, or there may be altered metabolism of the drug (Hyrich *et al.* 2006).

1.3.9. **Serum and gene expression biomarkers**

A recent study has identified the myeloid related protein 8/14 (MRP8/14) protein complex as a serum biomarker of treatment response to biologics in RA (Choi *et al.* 2013). This study found that in adalimumab (n=86), infliximab (n=60) and rituximab (n=24) pre-treatment serum of responders contained much higher levels of MRP8/14 (p<0.05). Gene expression studies hold the potential to identify predictors of treatment response in RA, however none have been identified to date (Smith *et al.* 2013).
1.3.10. Genetic factors as indicators of response to anti-TNF therapy

Response to several drugs is influenced by genetic variation. For example, the time and dose of warfarin, an anticoagulant, required to achieve an INR (a measure of the degree of anticoagulation) is under partial genetic control with both the *VKORC1* and *CYP2C9* genes both associated (Limdi *et al.* 2009; Natarajan *et al.* 2013). However, it is not clear whether genetic variation would be expected to influence response to biologic therapies in RA, and twin and family studies are not feasible to establish a genetic component. Methods have now been developed to allow heritability to be calculated from genome-wide SNP data; using two linear mixed-modelling approaches (Bayz and GCTA), a recent study has estimated that the heritability of treatment response to anti-TNF therapies is between 59-71% based on change in DAS28 (Umicevic *et al.* 2014). Therefore, there has been considerable interest in identifying genetic predictors of anti-TNF response as summarised below.
1.3.10.1.  **TNF genes**

It was speculated that genetic variants of *TNF* may influence responsiveness to anti-TNF therapies and five SNPs have been explored in detail as they capture much of the haplotypic variation at the *TNF* gene locus; these SNPs are located at (relative to the transcription start site) -308, -238 and -857 in the promoter genes and -676 and -196 in the receptor genes. Of these, -308 has been studied in the most depth as a possible genetic marker of response to anti-TNFs (Emery and Dorner 2011) the results of which are summarised in Table 1.3 below. Several of these studies, including two meta analyses involving 311 and 692 patients, found that the genotype *TNF*-308GG was associated with better response to anti-TNFs generally (OR= 0.33 and 0.43 respectively) (Lee *et al*. 2006; O’Rielly *et al*. 2009; Seitz *et al*. 2007). However, conflicting results have been found by more recent better powered studies that showed no significant association between response and *TNF*-308 genotype (Marotte *et al*. 2008; Maxwell *et al*. 2008; Ongaro *et al*. 2008; Swierkot *et al*. 2014). Such studies are hindered by the rare occurrence of the *TNF*-308AA genotype and heterogeneity in the genetic background; differences in study design also make comparison of the studies difficult (Emery and Dorner 2011). This shows further, better powered research into the association between *TNF*-308 genotype and response to TNF antagonists is necessary to conclusively decipher whether it could be used as a predictive marker for response.

A recent study involving 280 RA patients also investigated the *TNF*-857 variant and found it to be significantly associated with anti-TNF response (p= 0.045), but the *TNF*-238 variant was not significantly associated (Swierkot *et al*. 2014).
<table>
<thead>
<tr>
<th>Drug</th>
<th>TNF-308 genotype</th>
<th>Outcome measure</th>
<th>Results</th>
<th>Conclusion</th>
</tr>
</thead>
<tbody>
<tr>
<td>Etanercept</td>
<td>GG</td>
<td>Mean change in DAS28 score at 6 months</td>
<td>-2.51</td>
<td>AA genotype associated with poor response (p=0.001, n=7)</td>
</tr>
<tr>
<td>(Maxwell et al. 2008)</td>
<td>GA</td>
<td></td>
<td>-2.74</td>
<td></td>
</tr>
<tr>
<td></td>
<td>AA</td>
<td></td>
<td>-0.61</td>
<td></td>
</tr>
<tr>
<td>Etanercept</td>
<td>GG</td>
<td>Mean change in DAS28 score at 6 months</td>
<td>-2.23</td>
<td>GG genotype associated with good response (p=0.005, n=38)</td>
</tr>
<tr>
<td>(Guis et al. 2007)</td>
<td>GA</td>
<td></td>
<td>-1.69</td>
<td></td>
</tr>
<tr>
<td>Infliximab</td>
<td>GG</td>
<td>Mean change in DAS28 score at 6 months</td>
<td>-2.28</td>
<td>No significant association between genotype and response (p=0.8, n=17)</td>
</tr>
<tr>
<td>(Maxwell et al. 2008)</td>
<td>GA</td>
<td></td>
<td>-2.61</td>
<td></td>
</tr>
<tr>
<td></td>
<td>AA</td>
<td></td>
<td>-2.28</td>
<td></td>
</tr>
<tr>
<td>Infliximab</td>
<td>GG</td>
<td>Mean change in DAS28 score at 6 months</td>
<td>-2.29</td>
<td>GG genotype associated with good response (p=0.0086, n=41)</td>
</tr>
<tr>
<td>(Mugnier et al. 2003)</td>
<td>AA/GA</td>
<td></td>
<td>-1.24</td>
<td></td>
</tr>
<tr>
<td>Adalimumab</td>
<td>GG</td>
<td>Mean change in DAS28 score at 6 months</td>
<td>-2.5</td>
<td>GG genotype associated with good response (p&lt;0.043)</td>
</tr>
<tr>
<td>(Cuchacovich et al. 2006)</td>
<td>GA</td>
<td></td>
<td>-1.8</td>
<td></td>
</tr>
<tr>
<td>Infliximab</td>
<td>GG</td>
<td>Proportion of patients with ACR20 at 6 months (n=34), %</td>
<td>73</td>
<td>No significant association between genotype and response</td>
</tr>
<tr>
<td></td>
<td>GA</td>
<td></td>
<td>18</td>
<td></td>
</tr>
<tr>
<td></td>
<td>AA</td>
<td></td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>Etanercept</td>
<td>GG</td>
<td>Proportion of patients with ACR50 at 6 months (n=39), %</td>
<td>80</td>
<td></td>
</tr>
<tr>
<td></td>
<td>GA</td>
<td></td>
<td>18</td>
<td></td>
</tr>
<tr>
<td></td>
<td>AA</td>
<td></td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>Adalimumab</td>
<td>GG</td>
<td>Proportion of patients with ACR70 at 6 months (n=45), %</td>
<td>82</td>
<td></td>
</tr>
<tr>
<td>(Ongaro et al. 2008)</td>
<td>GA</td>
<td></td>
<td>15</td>
<td></td>
</tr>
<tr>
<td></td>
<td>AA</td>
<td></td>
<td>2</td>
<td></td>
</tr>
</tbody>
</table>

Table 1.3: Summary of the results of studies of TNF-308 promoter genotypes and their affect on response to specific anti-TNF drugs (adapted from Emery and Dorner 2011).
1.3.10.2. **TNF receptor genes**

Several TNF receptor (TNFR) genes polymorphisms have been investigated in a number of studies, with some reporting association of the TNFR1A-36 variant (Swierkot et al. 2014); a polymorphism in exon 6 of the TNFRII gene (Fabris et al. 2002); a variant 676bp from the TNFRII transcription start site (Ongaro et al. 2008) and polymorphism at -196 from the TNFRII transcription start site (Rooryck et al. 2008). However, the largest study, to date, which tested association across the whole of the TNFRII gene in 909 anti-TNF treated patients failed to corroborate these findings (Potter et al. 2010).

1.3.10.3. **Fc gamma receptor type IIIA**

Polymorphisms in the Fc gamma receptor type IIIA gene have been correlated with treatment response to anti-TNFs (Tutuncu et al. 2005). Analysis of the polymorphism at base pair 158 showed that patients who were homozygous for the low affinity phenylalanine containing allele tended to have a good clinical response to anti-TNFs (Tutuncu et al. 2005).

1.3.10.4. **RA susceptibility genes**

When looking for a genetic predictor of response to anti TNFs, another logical step is to prioritise the common susceptibility genes of RA. A study of 1334 patients which took this approach (Tan et al. 2010) analysed 18 SNPs which had previously been associated with RA susceptibility and found two of these were associated with good response to TNF antagonists. One of these SNPs was in the AFF3 gene located on chromosome 2q, where a G allele at the rs10865035 locus was associated with an improved clinical outcome (p=0.015). The other SNP associated with treatment response in the study (rs763361) was in the CD226 gene on 8q22. A
C allele at this locus was associated with poor response to anti-TNF therapy, measured using the EULAR response criteria (p=0.048). While both associations were statistically significant, their overall effect was modest (Tan et al. 2010).

In a study of 1283 RA patients which investigated 31 RA-associated SNPs, a SNP at the protein tyrosine phosphatase receptor type C (PTPRC) susceptibility locus (rs10919563) was tested for association with treatment response in good responders compared to nonresponders to etanercept, infliximab and adalimumab. This study found that the minor allele (A) was associated with poor treatment response (Cui et al. 2010). This was replicated more recently in 1,115 RA patients who were considered either good or poor responders according to the EULAR response criteria (Plant et al. 2012). However, this association was not replicated by a study from the CORRONA registry in the United States (Pappas et al. 2013).

PTPRC, also known as the CD45 antigen, is important in the regulation of TNF signalling which may provide an explanation of how a polymorphism in this gene affects response to anti-TNF drugs (Plant et al. 2012).
1.3.10.5. **Genes involved in immune response**

The NLRP3-inflammasome has been previously associated with inflammatory disorders, and recently a candidate gene study investigating 34 SNPs in the BRAGGSS cohort involving 1278 patients identified SNPs in the **CARD8** gene which were associated with response to anti-TNFs ($p=0.02$) (Mathews et al. 2014). Interestingly, the same study found that expression of that gene was higher in good responders, indicating the genetic variant could be altering gene expression.

The **CD84** gene, which encodes a leukocyte differentiation antigen, is associated with anti-TNF response at almost genome-wide significance levels. In a GWAS of over 2 million common variants in 2706 European RA patients, a SNP (rs6427528) in this gene was associated with response to etanercept ($p=8E-8$), but not adalimumab or infliximab (Cui et al. 2013). This variant is predicted to disrupt a transcription factor binding motif in the gene. The same study showed that expression of the gene was higher in RA and healthy individuals with the genotype associated with good response. Interestingly these findings were not replicated in a Japanese population, and a replication study in patients of Portuguese ancestry showed the same direction of effect though it was not significant (Cui et al. 2013).

In a candidate gene study of 187 SNPs in 909 patients, eight SNPs in the Toll-like receptors (TLRs) and nuclear factor kappa-light-chain-enhancer of activated B cells (NFκB) systems were associated with anti-TNF response ($p<0.05$) (Potter et al. 2010). These genes are important in the regulation of inflammation and immune responses. While these SNPs were associated with treatment response to anti-TNFs, replication experiments are still needed before this data can be confirmed (Potter et al. 2010).

The P38 signalling network was identified as a candidate for studies into genetic contribution to response due to its importance in inflammatory pathways, including
influence cytokine production (Coulthard et al. 2011). A study investigating 38 SNPs in 1102 RA patients identified associations between anti-TNF response and genes encoding proteins in the p38 signalling network. Seven SNPs in five genes were identified which were linked with improved therapy response measured using absolute change in DAS28 score and the EULAR response criteria. These genes showed an association with infliximab and adalimumab, though further investigation is necessary to determine if this association can be validated (Coulthard et al. 2011).

1.3.10.6. Novel genes

The only locus associated with anti-TNF response at genome-wide significance maps to PDE3A-SLCO1C1, which encodes a phosphodiesterase and a solute carrier organic anion transporter. The SNP (rs3794271), which maps to the gene body of SLCO1C1, was associated with response to etanercept, infliximab and adalimumab (p=1.74E-5) in a GWAS replication study of 315 Spanish RA patients. When this data was combined with the original GWAS data from a Danish population, the association reached genome-wide significance (p=3.3E-10) (Acosta-Colman et al. 2013).

A GWAS from the BRAGGSS cohort involving 566 RA patients in an initial study and two replication cohorts of 379 and 341 patients, identified seven SNPs at novel genetic loci with possible evidence for association with drug response in anti-TNFs, however subsequent studies have not replicated these findings (Plant et al. 2011). One of these SNPs maps to the PDZ domain containing protein 2 (PDZD2) gene (p=7.37E-4) which influences insulin secretion. This is an interesting association as elevated insulin levels are a feature of severe early RA (Plant et al. 2011). The study also identified a SNP associated with improved response which mapped to
the eyes absent homolog 4 (EYA4) gene (p=5.67E-5). Because of the role of EYA4 in expression of interferon-β, the association could be due to a possible link between DNA induced innate immune response and the efficacy of TNF antagonists (Plant et al. 2011).

1.3.10.7. Summary of current knowledge of genetic predictors of drug response

Recent studies have identified genetic markers associated with response at genome-wide significance, however these markers do not have sufficient predictive ability to influence clinical allocation of anti-TNFs. The studies described are still limited by sample size due to the rarity of patients who meet the entry criteria, ie patients with RA who are on biologic treatments. This also means it is difficult to have a homogenous study population, thus there are often underlying genetic differences between patients which may also act as confounders. Therefore, larger studies and meta-analyses are still necessary to identify genes which have a role in response to anti-TNFs.
1.3.10.8. **Missing heritability**

While genetic studies have provided a great increase in our knowledge of RA susceptibility and pathogenesis, there is still a substantial proportion of missing heritability in the disease, with existing genetic susceptibility loci explaining just 17% (Figure 1.5). Few genetic predictors of anti-TNF therapy have been identified and even fewer replicated, although heritability estimates suggest that there is a genetic component to treatment response. It is possible that some of this missing heritability is explained by rare genetic variants with small effect sizes, gene-gene interactions and epigenetic factors such as chromatin and DNA modifications, the most commonly studied of which is DNA methylation, investigated in this thesis.

Epigenetics is an emerging field of research which can be influenced by both genetic and environmental factors, which may explain some of the missing heritability of RA and treatment response and could also provide a possible mechanism by which environmental factors could influence these phenotypes. There is a growing body of evidence of to support an epigenetic component in both RA development and treatment response, which is explored in more detail in section 1.4.7.
1.4. Epigenetics

Epigenetics is a broad term that describes heritable features which alter gene expression and have a phenotypic effect without altering the DNA sequence (Gomez and Ingelman-Sundberg 2009). Epigenetic phenomena include DNA methylation, histone modification (which in turn affects chromatin structure) and regulatory non-coding RNAs such as microRNAs (Gomez and Ingelman-Sundberg 2009). There has been huge advancement in epigenetic research in recent years, resulting in the implication of epigenetic factors in the causation of many common diseases (Feinberg 2010). A huge advance in epigenetic research was the release of publically available resources which map the epigenome, such as the Encyclopedia of DNA Elements (ENCODE) and the Epigenome Roadmap. ENCODE (www.encodeproject.org) was produced through international collaborations that mapped whole genome transcription, binding of transcription factors, histone modifications and consequent chromatin structure (The ENCODE Project Consortium 2012). This allowed characterisation of the function of genetic elements in 80% of the genome and identification of regulatory elements. The Epigenome Roadmap (www.roadmapepigenomics.org) was set up by the National Institute for Health (NIH) which employs next generation technologies for sequencing of whole genome DNA methylation, RNA sequencing, histone modifications and chromatin structure (Bernstein et al. 2010). These resources have characterised epigenetic modifications in multiple cell types and tissues and improved the understanding of functional implications of genetic alterations as well as giving a greater understanding of the position and interplay of various genetic elements.
1.4.1. Histone modifications

Post-translational modification of the histone tails through acetylation/deacetylation, methylation, phosphorylation, ubiquitination and sumoylation can alter chromatin conformation, consequently affecting gene regulation (Strahl and Allis 2000). Acetylation of histones H3 and H4 are well studied, and are known to activate genes, while deacetylation is associated with gene repression (Wang et al. 2009). The balance of these is controlled by histone acetylase transferases (HATs) and histone deacetylases (HDACs) (Wang et al. 2009). Interestingly, HDAC inhibitors have anti-inflammatory activity and have been found to have therapeutic potential in RA (Grabiec et al. 2010; Joosten et al. 2011; Lin et al. 2007; Nasu et al. 2008; Nishida et al. 2004; Saouaf et al. 2009) however these inhibitors can also modify non-histone proteins, potentially limiting their widespread use (Grabiec et al. 2010).

Lysine and arginine residues in histone tails can be methylated, with the addition of one, two or three methyl groups (Strahl and Allis 2000). The amount and position of methylation on the histone tail determines whether it has an activating or repressing effect, for example H3K4me1 is associated with activation of a gene, while H3K27me3 is associated with repression of a gene (Shlyueva, Stampfel, and Stark 2014). Methylation of histones is regulated by histone methyltransferases. Interestingly, the histone methyltransferase enhancer of zeste homologue 2 (EZH2), which silences genes through methylation of H3K27, has been found to be overexpressed in RA synovial fibroblasts (Margueron and Reinberg 2011).
1.4.2. Non-coding RNA

Traditionally, research has been focussed on coding transcripts; however, in recent years the importance of non-coding RNA in regulation of genes has become apparent (Amaral et al. 2008; Ponting, Oliver, and Reik 2009). MicroRNAs (miRNAs) are small non-coding RNA molecules which are around 18-22 nucleotides in length and can regulate gene expression post-transcription (Bartel 2004). Interestingly, miRNAs have been linked with RA in studies of synovial fibroblasts (Niederer et al. 2012; Stanczyk et al. 2011) and more recently, in patient serum where different levels of miRNA-223 and miRNA-16 transcripts were identified in serum of individuals with early RA (n=34) compared to healthy controls (n=16) (Filkova et al. 2014). While the majority of research in non-coding RNAs has focussed on small regulatory RNAs, there is increasing interest in long non-coding RNAs (lncRNAs) which are structurally similar to miRNAs but are over 200bp in length (Birney et al. 2007; Rinn and Chang 2012). LncRNAs are being increasingly investigated in studies of disease, for example a very recent study found that methotrexate has a therapeutic effect in RA by increasing levels of a lncRNA (lincRNA-p21), which inhibits the activity of NF-κB (Spurlock, III et al. 2014).
1.4.3. DNA methylation

DNA methylation is the most widely studied epigenetic phenomena, and involves the attachment of methyl groups to C-G dinucleotides (CpG), creating 5-methylcytosine (5mC) residues (Figure 1.8) (van der Maarel 2008). These CpGs are present throughout the genome, with 70-80% of all CpGs being methylated (Bird 2002). CpGs also tend to cluster in CpG islands which usually lie at promoter regions of genes, which are usually hypomethylated (van der Maarel 2008). Methylation is known to influence gene expression thus it has been postulated that aberrant methylation could have a role in disease pathogenesis (van der Maarel 2008). This is supported by the finding that differentially methylated regions (DMRs) often overlap with disease-associated SNPs (Ziller et al. 2013). Changes in methylation of DNA have been found in almost all forms of cancer, where there is usually hypermethylation of tumour suppressor genes preventing their expression while global hypomethylation of DNA causes chromosome instability and increased gene expression. This transition fuels the characteristic uncontrolled division of cancer cells (Das and Singal 2004). DNA methylation can be altered by environmental factors, the most notable of which identified to date is smoking; for example, methylation of the F2RL3 gene has consistently been found to be altered by smoking (Breitling et al. 2011; Zhang et al. 2014).

1.4.3.1. Effect of DNA methylation on gene expression

The relationship between DNA methylation and gene expression is not entirely understood, and controversy remains as to whether it is a causal relationship, and if so, which direction that relationship is in. The characteristic methylation distribution of highly or lowly regulated genes is evidence of an association between the two phenomena (Ball et al. 2009). Dynamic methylation sites also correlate with gene regulatory elements, particularly enhancers and the binding sites of transcription
factors (Ziller et al. 2013). The simplistic model of this relationship assumes that methylation in promoter sequences of a gene prevent gene transcription, possibly through physical blockade of transcription factors, while unmethylated genes are transcribed (Figure 1.6), however recent studies have refined this theory and identified a complex relationship between methylation and gene expression that extends far beyond the promoters of genes.

A study from George Church’s group, based in Harvard, employed next generation sequencing technology to investigate methylation and its correlation with gene expression at around 7000 CpG sites in B lymphocytes, fibroblasts and induced pluripotent stem cells (Ball et al. 2009). This study showed that genes with different levels of expression had characteristic methylation profiles in different regulatory elements (Figure 1.7). The study showed that gene body methylation tended to be much higher in genes with a high level of expression, and this relationship was seen across the majority of the gene and in surrounding regions. The transcription start site however, displays the inverse relationship with highly expressed genes generally having very low methylation levels compared to lowly expressed genes (Figure 1.7). The correlation of expression levels with promoter methylation is most pronounced in promoters which have an intermediate number of CpG sites within the promoter region, which accounts for around 16% of promoters. Promoters with a large number of CpG sites were found to generally have low methylation while those with a small number of CpG sites tended to have high methylation, regardless of expression levels in the genes (Ball et al. 2009).

The variability of methylation sites in specific gene elements is also informative, and a recent study in which genome-wide bisulphite sequencing was performed on 42 samples in 30 tissue types, found that enhancers are one of the most dynamic genetic elements, with 26% of DMRs mapping to enhancer regions (Ziller et al. 2013). Due to the bias of current methylation array-based platforms towards
focussing on promoter methylation (Bibikova et al. 2011), this indicates that bisulphite sequencing of enhancer regions could be important in future studies of methylation. Methylation variable regions identified at gene regulatory elements have also led to the identification of key regulators of cell lineage specificity (Ziller et al. 2013).

Methylation is generally associated with gene repression, which is thought to happen when the methylation marks recruit methylation binding proteins, which in turn recruit histone deacetylases (HDACs) which cause chromatin remodelling, consequently repressing expression of the associated gene (Bogdanovic and Veenstra 2009; Martinowich et al. 2003). There are 5 prolific methylation binding proteins, each of which contains a methyl-CpG-binding domain (MBD); Methyl-CpG-binding protein 2 (MeCP2), MBD1, MBD2, MBD3 and MBD4. The most important methylation binding protein is MeCP2, which interacts with a wide range of proteins which have a role in transcriptional repression including HDACs (Nan et al. 1998; Rietveld, Caldenhoven, and Stunnenberg 2002) and histone methyltransferases (Fuks et al. 2003). Interestingly the MeCP2 protein also interacts with the DNA methyltransferase enzyme 1 (DNMT1), indicating it also has a role in the targeting of maintenance of DNA methylation (Kimura and Shiota 2003).

While generally methylation is associated with repression of gene expression, recent proteomic and bisulphite sequencing studies have shown that transcription factors can interact with methylated DNA, and certain genes with methylated promoters are highly transcribed (particularly in pluripotent cells), indicating DNA methylation can have an alternative role in gene activation (Spruijt and Vermeulen 2014).
Figure 1.6: Simplistic postulated mechanism by which DNA methylation in a promoter can influence transcription of a gene. Methylation in the promoter region of the gene is generally thought to prevent transcription of a gene, possibly by physical blockade of a transcription factor binding site. When the promoter is unmethylated, transcription factors are free to bind and hence the gene can be activated.
Figure 1.7: Methylation levels in regions relative to a gene have expression-dependent profiles. High gene body methylation is seen in highly expressed genes when samples are interrogated with bisulphite padlock probe based sequencing (a) and methyl-sensitive cut counting methylation profiling (b). The inverse relationship is seen at the transcription start site, and seems to be continuous (c). Expression related differences in DNA methylation levels are also associated with distance from the transcriptional end point of a gene (d). (Ball et al. 2009)
1.4.3.2. *Maintenance of methylation*

Methylation throughout the mammalian genome is initiated and maintained by DNMTs (van der Maarel 2008). During development, the majority of methylation is established by the DNMT3a and DNMT3b DNA methyltransferases during early embryogenesis (Kafri *et al.* 1992; Monk, Boubelik, and Lehnert 1987). Following implantation, the primordial germ cells undergo further methylation changes, with reestablishment of methylation following almost complete demethylation (Cedar and Bergman 2009; Cheng and Blumenthal 2008; Reik 2007). Because there is not complete erasure of methylation between generations, the possibility of transgenerational inheritance of methylation has been postulated and supported by studies such as the Dutch famine study (Heijmans 2014). However, the validity of the evidence in favour of transgenerational inheritance has been widely questioned, in part due to the genetic component of methylation heritability (see section 1.4.4 for more detail) making transgenerational epigenetic inheritance a hotly debated topic within the epigenetic community, with consensus yet to be achieved (as discussed in (Heard and Martienssen 2014) and (Grossniklaus *et al.* 2013)).
1.4.3.3. **5-hydroxymethylation**

Recently, another modification of cytosine residues (Figure 1.8) called 5-hydroxymethylcytosine (5hmC) has been found to be more widespread than initially thought, particularly in brain tissue with 13% of all CpGs modified by hmC in the adult brain (Wen et al. 2014). This modification has also been implicated in diseases including pediatric brain tumors and uterine leiomyoma (Ahsan et al. 2014; Mariani et al. 2014; Navarro et al. 2014) and in development (Bakhtari and Ross 2014). An emerging issue is that during bisulphite conversion, which is currently a widespread technique to differentiate methylated and unmethylated CpG sites (see section 2.2.4.1 for more detail), methylated and hydroxymethylated cytosine residues are indistinguishable, however methods are being developed to combat this (Booth et al. 2012; Stewart et al. 2014).

![Chemical structure of cytosine nucleotides](image)

**Figure 1.8:** Chemical structure of modified and unmodified cytosine nucleotides. The chemical structure of cytosine (C), 5-methylcytosine (5mC) and 5-hydroxymethylcytosine (5hmC) are shown. Modifications are covalently attached to the fifth carbon molecule in the cytosine base, and can modify the behaviour of the cytosine residue. Both 5mC and 5hmC behave the same when exposed to sodium bisulphite.
1.4.4. Heritability of DNA methylation

Twin studies of DNA methylation have been important in establishing heritability of this phenomenon. The largest twin study to date investigated DNA methylation in adipose tissue of 648 twins including 97 monozygotic (MZ) pairs, 162 dizygotic (DZ) pairs and 130 singletons. This study found that methylation variation was highly heritable, with 28% of CpG sites associating with nearby SNPs (Grundberg et al. 2013). This is in contrast to previous findings, which tended to indicate DNA methylation has low heritability. A study focussing on methylation variability at 1760 CpG sites in the MHC region in CD4+ lymphocytes in 49 MZ and 40 DZ twin pairs found that heritability of methylation was between 2 and 16%, and varied in different types of genetic region (eg CGIs, 5’ ends of genes, conserved noncoding regions and randomly selected regions) (Gervin et al. 2011). This study suggested that there is a genetic influence on DNA methylation, as MZ twins were more similar to each other than DZ twins, however it was concluded that the majority of variation arises from non-genetic (environmental) factors (Gervin et al. 2011). This was supported by evidence from another small study of 22 MZ and 12 DZ twin pairs which found MZ pairs had lower discordance of methylation at birth; the authors suggested that the majority of variability could be attributed to intrauterine environment, birth weight and stochastic factors (Gordon et al. 2012). Another small study in buccal cells found that methylation in CpG islands was more similar within MZ pairs compared to CpG island (CGI) shores, shelves and other non-CGI regions (van et al. 2014). They found that this effect was more pronounced in hypomethylated sites. Recently, a non-twin based family study of genome-wide methylation heritability involving 614 individuals concluded that genetic variation was controlling transgenerational methylation similarity (McRae et al. 2014).

Studies have also identified locus-specific differences in heritability of methylation, for example the previously described twin study focussing on the MHC region found
larger within-pair differences in the conserved noncoding and randomly selected regions compared to promoter regions (Gervin et al. 2011). This is supported by the study of neonatal twins which found that CpG islands, which are predominantly in gene promoters, had the lowest within-pair differences, and this effect decreased with distance from the island (Gordon et al. 2012). Another study also found that CpG island and promoters were the least variable genomic elements within-pairs, and interestingly they found that this effect was tissue-specific, as it was apparent in white blood cells, but not buccal epithelial cells (Kaminsky et al. 2009). The largest twin study to date also identified promoters as the genomic elements with the lowest variability within pairs (Grundberg et al. 2013).
1.4.5. DNA methylation research techniques

The study of methylation in DNA and epigenetics generally is a relatively new area of scientific research and yet the research techniques are developing at an extraordinary rate (Laird 2010). Initial research was restricted to analysis of specific loci, while more recent technologies allow analysis of the entire methylome at single base resolution (Laird 2010). There is a plethora of techniques available which take different approaches, with differences in pre-treatment of DNA and analytical steps. The main methods of pre-treatment of DNA used are digestion with endonucleases, affinity enrichment and sodium bisulphite (Laird 2010). The analytical steps used include locus specific analysis, gel based analysis, array based analysis and next generation sequencing (NGS) based analysis (Laird 2010). A summary of the techniques which employ each of these strategies is shown below in Table 1.4 (Laird 2010).
<table>
<thead>
<tr>
<th>Pre-treatment</th>
<th>Analytical step</th>
<th>Locus-specific analysis</th>
<th>Gel-based analysis</th>
<th>Array-based analysis</th>
<th>NGS-based analysis</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Enzyme digestion</strong></td>
<td></td>
<td><em>HpaII-PCR</em></td>
<td>Southern blot</td>
<td>DMH</td>
<td>Methyl-seq</td>
</tr>
<tr>
<td></td>
<td></td>
<td>RLGS</td>
<td>MCAM</td>
<td>HELP</td>
<td>MCA-seq</td>
</tr>
<tr>
<td></td>
<td></td>
<td>MS-AP-PCR</td>
<td>HELP</td>
<td>HELP-seq</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>AIMS</td>
<td>MethylScope</td>
<td>MASS</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>CHARM</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Affinity enrichment</strong></td>
<td></td>
<td>MeDIP-PCR</td>
<td>MeDIP</td>
<td>MeDIP-seq</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>mDIP</td>
<td>MIRA-seq</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>mCIP</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>MIRA</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Bisulphite conversion</strong></td>
<td></td>
<td>MethyLight</td>
<td>Sanger BS</td>
<td>BiMP</td>
<td>RRBA</td>
</tr>
<tr>
<td></td>
<td></td>
<td>EpiTYPER</td>
<td>MSP</td>
<td>GoldenGate</td>
<td>BC-seq</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Pyrosequencing</td>
<td>MS-SNuPE</td>
<td>Infinium</td>
<td>BSPP</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>COBRA</td>
<td></td>
<td>WGSBS</td>
</tr>
</tbody>
</table>

Table 1.4: Techniques which employ different strategies for methylation analysis of DNA (adapted from Laird 2010).
1.4.5.1. *Digestion with endonucleases*

A map of methylated sites in DNA can be created by examining the DNA fragments produced after treatment with methylation sensitive endonucleases (Laird 2010). This approach involves the initial methylation dependent digestion of the DNA followed by polymerase chain reaction (PCR). This approach can be used for genome wide analysis in techniques such as restriction landmark genome scanning (RLGS), array based technologies such as methylated CpG island amplification (MCA) and differential methylation hybridisation (DMH) in which methylation-sensitive enzymes digest a pool of DNA while another pool is mock digested (Laird 2010). Sequencing approaches which use restriction enzymes are more powerful than the use of enzymes in arrays as they allow analysis of specific alleles (Laird 2010).

1.4.5.2. *Affinity enrichment*

Affinity enrichment takes a different approach to identifying methylated DNA, as it uses specific antibodies or methyl binding proteins to give a methylation profile of the DNA. This is a very quick technique, though has low resolution and is not good for identifying individual methylation variable positions (MVPs). It is often used in combination with array hybridisation (Laird 2010).
1.4.5.3. **Methylation analysis by sequencing**

Sanger sequencing was the first widely used sequencing method; first described over 35 years ago, it has been developed and improved vastly since then (Sanger, Nicklen, and Coulson 1977). This technique relies on dideoxy chain termination. This involves PCR amplification of a sample in the presence of dideoxy nucleotides which, due to random incorporation through the amplification, eventually terminate at each respective base in the sequence. Through polyacrylamide separation of samples, the sequence can be determined by the comparative position of each band. This technique can be employed for methylation analysis following the bisulphite conversion of a DNA sample. While double peaks at a particular locus would traditionally indicate heterozygotes, this can instead be interpreted as the degree of methylation at a CpG site, eg if a pure ‘T’ peak is produced, this indicates complete bisulphite conversion therefore an unmethylated locus, while a mixed ‘T’ and ‘C’ peak would indicate partial methylation (or incomplete bisulphite conversion) (Frommer et al. 1992).

Pyrosequencing technology relies on a sequence-by-synthesis method whereby primers are used to direct polymerase extension using deoxyribonucleotide triphosphates (dNTPs) (Tost and Gut 2007). For each dNTP added to the sequence, a pyrophosphate is released. Enzymatic conversion of this pyrophosphate molecule to adenosine triphosphate (ATP) then drives the conversion of luciferin to oxyluciferin. This conversion produces a light signal which is directly proportional to the amount of dNTP initially incorporated, thus detection of this light signal by a charge-coupled device chip allows the accurate quantification of dNTPs incorporated at a particular base location (Tost and Gut 2007). This technique was originally designed for sequencing genomic DNA in 1996 (Ronaghi et al. 1996) but has since been developed and improved to a much cheaper and more accurate technology. It has also been adapted for methylation analysis, and addition
of a methylation module to the software design means this technique has become a widely used technique for quantitative analysis of methylation status of CpG sites at a single base pair resolution (Tost and Gut 2007). For a more detailed description of pyrosequencing technology please see section 2.2.6.1.

In recent years, whole genome bisulphite sequencing has become increasingly common, though it is still considered a prohibitively expensive technique for most studies. A study which mapped the entire methylome in 30 different tissue types concluded that for many studies, whole genome bisulphite sequencing is inefficient and unlikely to yield interesting results, and suggested reduced representation bisulphite sequencing approaches would capture the majority of informative methylation sites by targeting DMRs that were found to be most variable (Ziller et al. 2013).
1.4.5.4. **Genome-wide methylation profiling**

Bisulphite conversion has allowed the development of techniques which allow comprehensive coverage of the human genome with relatively low labour intensity (Laird 2010). The approach which has made most use of bisulphite conversion to identify methylation is array hybridisation (Laird 2010). Several arrays have been designed, with increasing genome coverage and resolution (Laird 2010). However, the main problem with array hybridisation using bisulphite converted DNA is the reduced sequence complexity due to the conversion of unmethylated cytosine residues being converted to thymine residues, which could lead to decreased hybridisation specificity, thus during the design of arrays this must be taken into account when designing the probes (Laird 2010).

The Illumina GoldenGate Bead Array uses multiplexed primer extension with fluorescently labelled primers that are specific to methylated and unmethylated sequences at each site (Bibikova *et al.* 2006b; Bibikova *et al.* 2006a; Bibikova and Fan 2009). The array covers 1536 CpG sites in 807 genes. This array is specific for humans, and due to the genes which are targeted by the probes, it is often used in cancer research (Laird 2010).

The Illumina Infinium platform requires amplification of the whole bisulphite converted genome, which is then fragmented and hybridised to beads with methylation specific DNA oligomers (Bibikova *et al.* 2009). Each bead has a probe which corresponds to a certain methylation state at a specific CpG site (Bibikova *et al.* 2009; Brunner *et al.* 2009; Steemers and Gunderson 2007). There are currently two Infinium arrays available, the HumanMethylation27 DNA Analysis BeadChip and the HumanMethylation450 BeadChip DNA Analysis BeadChip. The original 27K version allowed the analysis of 27,578 CpG sites, including 14,495 protein coding genes (Bibikova *et al.* 2009) while the 450K version gives a more comprehensive coverage of more than 450,000 CpG sites (Dedeurwaerder *et al.* 2011). The
HumanMethylation450 BeadChip is currently the method of choice for the majority of epigenome-wide association studies (EWAS) due to its' coverage of CpG sites in the genome and relative inexpense.
1.4.6. **Considerations in DNA methylation study design**

When conducting investigations into DNA methylation, study design is of paramount importance in ensuring the reliability of the data produced, and in appropriate interpretation of the results. DNA methylation studies differ from genetic studies, as they can be directly confounded by environmental factors due to their influence on methylation levels. Due to the role genetic sequence plays in DNA methylation patterns (discussed in section 1.4.4), this is another important confounder which should be considered. Methylation study design must also consider the sensitivity of experimental techniques to batch effects, which will have a much greater impact due to the fact that methylation measurements are not discrete. The design of methylation studies is limited by the availability of appropriate sample collections, as many existing collections were intended for genetic studies, which had fewer potential confounders. In this section I will outline advantages and disadvantages of different study designs for the investigation of DNA methylation.
1.4.6.1. **Case control studies**

The majority of studies investigating DNA methylation in relation to disease compare cases with established disease with unrelated controls. An advantage of this study design is that large cohorts of this type are widely available making access to a large sample size relatively easy. The main disadvantage of this is that in such studies it is impossible to differentiate methylation differences which are causing disease, from those which are caused by disease pathogenesis or the effect of treatment (Rakyan et al. 2011b). Another disadvantage of such studies is that there are genetic differences in the comparison groups which could be driving apparent methylation differences, which may lead to overestimation of the direct effect of methylation on disease pathogenesis. This has particular impact if the technique used to measure methylation can be biased by genetic sequence, for example SNP-containing probes in array technologies can cause preferential binding which may make a genetic difference appear as a methylation difference (Price et al. 2013).

1.4.6.2. **Twin studies**

Studies of disease discordant MZ twins allow the identification of methylation differences without the confounding of genetic sequence (Bell and Spector 2011; Kaminsky et al. 2009). Investigation into methylation in MZ twin pairs also allows comparison of groups which are perfectly matched for age, gender and many environmental factors. Unfortunately very few sample collections of twins exist, and so many twin studies are underpowered (Rakyan et al. 2011b). Such studies are very rarely longitudinal, and so twin studies still face the limitation of identifying methylation differences which may be a cause or effect of disease. Despite these limitations, investigations into disease discordant monozygotic twins give a unique insight into methylation differences that do not have a genetic basis.
1.4.6.3. *Longitudinal studies*

Longitudinal cohorts have the potential to remove the fundamental issue in many DNA methylation studies of disease, by investigating methylation in pre-diagnosis samples of individuals who go on to develop disease and establishing the temporal origins of DNA methylation changes. This was demonstrated very well in a study of type-1 diabetes which identified methylation differences in twins with established disease, then showed these differences preceded disease diagnosis in samples from a longitudinal cohort (Rakyan *et al.* 2011a). Methylation differences identified in such studies are not a consequence of treatment, and are unlikely to be caused by disease pathogenesis, therefore are more likely to be causative of disease and could act as a pre-diagnosis marker of disease (Rakyan *et al.* 2011b).

In the investigation of treatment response, longitudinal cohorts are the gold standard for identification of DNA methylation biomarkers of response, as they allow investigation into pre-treatment samples, and comparison of pre- and post-treatment samples to identify methylation changes which are a consequence of therapy. While these differences may be due to evolving disease pathology, as individuals in such studies typically have established disease, they may still be informative as biomarkers.
1.4.6.4. **Tissue type**

In the design of epigenetic studies, it is important to consider that different cell types have distinct methylation signatures (Glossop *et al.* 2013; Reinius *et al.* 2012), which can confound studies if they are not performed appropriately. This was demonstrated recently in a study that found previously identified age-associated methylation signatures were actually driven by differences in cell composition (Jaffe and Irizarry 2014). These findings indicate that it is important to either investigate methylation in individual cell types, or to account for cellular heterogeneity during analysis (Aryee *et al.* 2014; Houseman *et al.* 2012). The revelation of the importance of cell type in methylation studies has led to an increase in cohorts in which specific cell types are isolated, or in which PBMCs are collected and cryopreserved with the intention of isolating individual cell types and performing methylation analysis in the future. While this is becoming more widespread, no study has yet investigated the effect of this cryopreservation on the DNA methylation. While cell composition is an important consideration, investigating individual cell types may not always be an option. For example, in the identification of clinically useful biomarkers of, for example, treatment response, a test that can be performed on whole blood would be preferable due to practical and economic constraints.
1.4.7. DNA methylation in rheumatoid arthritis

The heritability of RA in UK population estimated from a large twin study is 53% (95% CI 50-77) (MacGregor et al. 2000). This demonstrates the influence of both genetic and environmental factors in this complex disorder. While epigenetic mechanisms influence gene expression, they are also influenced by the environment, which suggests they may provide the mechanism by which genetic and environmental factors interact (Ballestar 2011). By analysing the DNA methylation in patients with RA and other rheumatic diseases, compared to healthy controls, several epigenetic studies have found a link between methylation and these disorders (Ballestar 2011).

Several case control studies have reported aberrant methylation in RA (Glossop et al. 2014; Kennedy et al. 2014; Nakano et al. 2013; Richardson et al. 1990). However, these studies cannot differentiate between cause and effect of the disease. They may also be subject to confounding due to differences in recruitment of cases and controls and biases due to case-control differences in the measurement of risk factors (Viatte, Plant, and Raychaudhuri 2013). The first methylation study in RA found reduced global methylation in T-cells from 10 RA patients compared to healthy controls (Richardson et al. 1990). This reduced 5-methylcytosine content has also been demonstrated in synovial fibroblast cells in RA patients (Corvetta et al. 1991).
1.4.7.1. **Targeted gene methylation studies**

Following the identification of global methylation changes, several studies were performed to investigate more targeted gene assays measuring methylation in RA. A small study involving just 6 RA patients identified a single unmethylated CpG motif in peripheral blood mononuclear cells of RA patients (Nile *et al.* 2008). The motif was in a regulatory region of the IL-6 gene, which the authors proposed as the mechanism by which IL-6 becomes highly expressed in RA pathogenesis (Nile *et al.* 2008). A study primarily focussing on systemic lupus erythematosus (SLE) found IL10 and IL1R2 genes to be hypomethylated in RA patients (Lin *et al.* 2011). Investigation into methylation of the TNF gene found that reduced methylation in the promoter region in myeloid lineage cell lines correlated with high expression levels of TNF (Sullivan *et al.* 2007). Due to the inflammatory role of TNF in RA pathogenesis, this suggests methylation may be involved in over production of TNF in RA (Sullivan *et al.* 2007).

A study which investigated expression and promoter methylation of the *CXCL12* gene in RA synovial fibroblasts (n=14) and OA synovial fibroblasts (n=11), found that there was lower promoter methylation in the RA samples, which correlated with higher expression of the gene (Karouzakis *et al.* 2011).

A very recent study identified a DMR upstream of the *FOXP3* promoter, which displayed enhancer activity which was silenced by methylation in RA Treg cells when compared to healthy controls (Kennedy *et al.* 2014). The RA Treg cells also displayed lower DNA methyltransferase expression and consequently had lower methylation in the DMR.
1.4.7.2. **Epigenome-wide association studies**

Recently epigenome-wide association studies (EWAS) of disease have become much more widespread and well powered, with the largest EWAS of RA involving 345 patients with RA and 337 unrelated controls (Liu et al. 2013). This study identified two clusters in the MHC region that mediated genetic risk for RA development. They also identified one non-MHC differentially methylated position (DMP) which mediated genetic risk, and which mapped to the promoter of the GSTA2 gene belonging to a supergene family that has previously been implicated in RA susceptibility in GWAS studies (Bohanec et al. 2009; Keenan et al. 2010; Lundstrom et al. 2011; Yun et al. 2005).

A consideration in epigenetic studies of disease is that, depending on study design, it is usually impossible to differentiate between epigenetic alterations which are cause or effect of disease. To account for this, Liu et al. performed causal inference testing to remove methylated sites which were likely to be consequential of disease, and focussed on sites which were mediating genetic associations.

Another problem epigenetic studies often face is cellular heterogeneity. Due to differences in methylation between cell types (Glossop et al. 2013; Reinius et al. 2012), apparent methylation differences identified in whole blood can actually be due to differences in cell composition. This was demonstrated recently in studies of ageing in which most of the apparent age-related methylation differences were actually due to cell composition (Jaffe and Irizarry 2014). The Liu et al. study was performed in whole blood, though they employed a statistical algorithm developed by Andres Houseman to estimate cell composition from the HumanMethylation450 BeadChip measurements of methylation using a reference dataset of separated cell types (Houseman et al. 2012; Reinius et al. 2012). These cell estimates were then used to adjust the dataset, reducing the effects of cell heterogeneity during the identification of DMPs, which they then validated using isolated monocyte fractions.
Following this adjustment, they saw a substantial decrease in significantly differentially methylated positions.

A recent replication study of the Liu 2013 findings found five of the ten DMPs identified were differentially methylated in the same direction in a small European cohort (n=46 RA patients, 15 healthy controls) and a small North American cohort (n=21 RA cases, 12 healthy controls) (van Steenbergen et al. 2014). One site in particular, cg21325723 in the MHC replicated very strongly, indicating it is the most reliable epigenetic marker of RA susceptibility identified to date. One of the sites that did not replicate was the non-MHC site mapping to the GSTA2 gene. While this replication study strengthens some of the findings from Liu et al, the small sample size is limiting. However a strength of the study is that the patients were recently diagnosed with RA so the methylation alterations are less likely to be consequence of disease pathogenesis (van Steenbergen et al. 2014).

Studies focussing on individual cell types to elucidate methylation differences in RA are becoming increasingly common though are still invariably underpowered. A recent EWAS in T and B lymphocytes identified 509 and 252 RA-associated DMPs respectively (Glossop et al. 2014). A subset of these DMPs (32 in T-lymphocytes and 20 in B-lymphocytes) showed larger changes in methylation, with a change in $\beta$ of 0.2 (which equates to approximately 20% difference in methylation at these sites). The small sample size of 12 cases with established seropositive RA and 12 controls means that these results should be treated with caution; however, because the investigation involves individual cell types, the power to detect methylation changes is increased due to the lack of cellular heterogeneity.

An EWAS using fibroblast-like synoviocytes (FLS) cell lines from RA patients (n=6) identified 207 differentially methylated genes when compared with FLS osteoarthritis (OA) lines (n=5) using the HumanMethylation450 BeadChip. All
participants in the study were female, however there was a difference in mean age (RA group=53 OA=68). Several of the genes identified have a previously confirmed role in RA including \textit{CHI3L1, CASP1, STAT3, MAP3K5, MEFV and WISP3} (Nakano \textit{et al.} 2013). A study from the same group found that in the RA FLS cells, DNMT function was reduced following exposure to the pro-inflammatory cytokine IL-1, causing reversible methylation alterations (Nakano, Boyle, and Firestein 2013). A study investigating genome-wide methylation using an alternative platform, the human 244K CpG island microarrays (Agilent, USA) in RA FLS (n=7) compared to OA (n=4) and healthy FLS (n=2) identified hypermethylation in the \textit{EBF3} and \textit{IRX1} genes, which are involved in TGF-\textit{\beta} signalling (Park \textit{et al.} 2013). These genes were also found to have reduced expression in RA FLS, indicating that the methylation may be controlling expression in the TGF-\textit{\beta} pathway.

While the majority of studies have supported aberrant methylation as having a role in RA pathogenesis, a small study which looked at several rheumatic diseases found no significant widespread changes in DNA methylation in monozygotic twins discordant for RA (Javierre \textit{et al.} 2010).

The body of research elucidating the role of DNA methylation in RA pathogenesis is rapidly increasing, though many studies remain underpowered and often confounded by age, gender and genetic factors. The identification of a genetic element in the inheritance of DNA methylation (Grundberg \textit{et al.} 2013), and the importance of genetics in RA susceptibility (Okada \textit{et al.} 2014) makes the consideration of underlying genetic sequence important in investigations into RA-associated DNA methylation. Studies of disease-discordant MZ twins are an ideal model to allow identification of DNA methylation differences in a group which is unconfounded by age, gender and genetic sequence.
1.4.8. DNA methylation in other inflammatory diseases

1.4.8.1. Systemic lupus erythematosus

A study investigating global methylation patterns of DNA samples derived from white blood cells in SLE discordant monozygotic twins and found hypermethylation in the SLE twins (Javierre et al. 2010). Further analysis including monozygotic and dizygotic twins identified 49 genes which were hypomethylated in the SLE twins compared to the healthy twins and matched controls, as shown in the heat map below (Figure 1.9) (Javierre et al. 2010). Some of these genes have previously been implicated in the pathogenesis of SLE strengthening a role for these genes in SLE development (Javierre et al. 2010). Another study also found that hypomethylation of DNA may be associated with SLE, specifically in the \textit{IL10} and \textit{IL1R2} genes (Lin et al. 2011). In this study, it was also found that reduced methylation of \textit{IL-10} and \textit{IL-1R2} was associated with more severe disease (Lin et al. 2011). A more recent EWAS study of CD4+ T cells in 18 SLE patients and 18 healthy controls identified 47 genes which contained DMPs (Coit et al. 2013). A large proportion of these genes were regulated by interferon.

SLE is characterised by reduced membrane CD5 expression. CD5 has two isoforms: CD5-E1A is expressed on the membrane and CD5-E1B is retained within the cell cytoplasm (Garaud et al. 2009). A recent study showed patients with SLE had reduced methylation and therefore up-regulation of the CD5-E1B isoform compared with healthy controls, consequently expression of the membrane isoform of CD5-E1A was reduced (Garaud et al. 2009). It was also found that high levels of IL6 in B cells of patients with SLE prevent the induction of DNA methyltransferase enzymes, thus reducing DNA methylation. This may explain the reduced methylation of \textit{CD5} genes in SLE (Garaud et al. 2009).
Figure 1.9: Heat map showing methylation of 49 genes identified as differentially methylated in 5 systemic lupus erythematosus discordant twin pairs and matched controls. Scale shows methylation status higher (red) and lower (blue) than average for each gene (Javierre et al. 2010).
1.4.8.2. **Type 1 diabetes**

A very well designed EWAS of type-1 diabetes (T1D) identified and validated methylation differences in twins then established the temporal origins of the changes using non-twin individuals (Rakyan *et al.* 2011a). The study identified 132 DMPs in CD14+ monocytes in T1D discordant twins (n=15 twin pairs), which they validated in 5 additional T1D discordant MZ pairs. To elucidate whether these DMPs were cause or effect of disease/treatment, they investigated methylation in CD14+ cells in pre-diagnosis samples and samples taken at diagnosis in 7 individuals. The study concluded that aberrant methylation in T1D happens very early in disease course and could act as a predictor of disease (Rakyan *et al.* 2011a).

1.4.8.3. **Type 2 diabetes**

A recent EWAS of adipose tissue in type 2 diabetes (T2D) discordant MZ twins (n=14 twin pairs) and T2D patients (n=120) versus unrelated controls identified no significant DMPs between twins, but found that 15,627 sites were differentially methylated between unrelated cases and controls at a false discovery rate (FDR) of 15% (Nilsson *et al.* 2014). The majority of these were of intermediate methylation values (between 20 and 70%) indicating that in T2D less-extreme methylation values are more likely to be variable. The identification of DMPs in unrelated case-control context but not disease discordant twins could be due to sample size, or due to genetic similarity between individuals (Nilsson *et al.* 2014). A targeted gene study in 93 newly diagnosed T2D patients found that 59% of the CpGs investigated in the promoter region of *TCF7L2*, the top susceptibility gene for T2D, were differentially methylated (Canivell *et al.* 2014). DMPs in T2D have been associated with prenatal exposures such as low maternal glucose, indicating that the methylation alterations in autoimmune disorders can be in place long before disease symptoms are apparent (Quilter *et al.* 2014).
1.4.8.4. **Sjogren’s syndrome**

Sjogren’s syndrome (SS) is a poorly understood autoimmune disorder characterised by dry mouth and eyes, and can develop on its own or alongside other common autoimmune conditions including SLE and RA. A recent EWAS of SS CD4+ T cells in 11 patients and 11 controls identified 311 hypomethylated genes and 115 hypermethylated genes associated with disease (Altorok et al. 2014). Interestingly, many of these genes were in the interferon signalling pathway, as found in SLE methylation studies (see section 1.4.8.1). Another gene which was differentially methylated was *RUNX1*, which has been associated with RA pathogenesis in Japanese populations (Tokuhiro et al. 2003). The overlap in the SS DMPs with those from SLE and with genetic associations in RA is particularly interesting due to the co-occurrence of the disorders, and could provide a potential shared mechanism which explains why the disorders often overlap.

Another study of 8 SS cases and 8 healthy controls investigated methylation levels in T-cells, B-cells and salivary gland epithelial cells. While the T and B cells showed no substantial differences in methylation, the salivary gland epithelial cells had a global decrease in methylation (Thabet et al. 2013). Interestingly, the study also found that this decrease in methylation was associated with a 7 fold decrease in DNMT1 expression (Thabet et al. 2013). This suggests that during disease, altered DNMT1 expression is causing aberrant demethylation of a range of other genes.
1.4.8.5. **Juvenile Idiopathic Arthritis**

The first study of JIA epigenetic dysregulation measured genome-wide methylation on the HumanMethylation27 BeadChip in CD4+ T cells from 14 JIA cases and healthy matched controls (Ellis et al. 2012). 145 DMPs were identified at a FDR of 10%, though the DMPs were different when patients were stratified for treatment with methotrexate. In treatment-naive individuals (n=10), just 11 DMPs were identified as statistically significant, however this may be due to reduced power caused by the decrease in sample size. One of the differentially methylated genes in the treatment naive cohort was the gene encoding inflammatory cytokine IL32 (Ellis et al. 2012).
1.4.9. **Epigenetic influences in therapy response**

Drug response in humans is controlled by very complex pathways which involve enzymes, transcription factors and receptors, many of which are under epigenetic control (Gomez and Ingelman-Sundberg 2009). For example, the cytochrome P450 (CYP) superfamily includes many genes, involved in the metabolism of drugs, which are controlled by methylation (Gomez and Ingelman-Sundberg 2009). This is supported by the observation that treatment with DNMT inhibitors results in an increase in expression of these genes (Nakajima, Iwanari, and Yokoi 2003). Another mechanism by which methylation influences drug response is by influencing the expression of transporter proteins, for example the human multidrug resistance gene 1 (MDR1) (Baker et al. 2005). MDR1 is overexpressed following treatment of breast cancer with certain chemotherapeutic drugs (Baker et al. 2005). This has been found to be caused by the interaction between drugs and DNA, in which MDR1 expression is only induced when the gene is hypomethylated (Baker et al. 2005).

Epigenetics also affects drug response by influencing expression of drug targets and receptors. For example, the expression of the chemokine receptor CXCR4 is regulated by methylation (Sato et al. 2005). This gene is involved in leukocyte trafficking and is a target for human immunodeficiency virus (HIV) therapy. CXCR4 is upregulated in several cancers and is seen as a possible new target for cancer treatment (Sato et al. 2005).
1.4.9.1. *Pharmacoeigenetics in cancer*

The role of epigenetics in drug metabolism and the expression of drug targets inevitably lead to the question of whether epigenetic status could be a predictor of response to particular drugs. The majority of studies addressing this question, to date, have focussed on cancer. The development of resistance to anti-cancer treatment has become a major problem in the treatment of malignancies (Glasspool, Teodoridis, and Brown 2006). Until recently, research to explain this resistance tended to involve searching for a single gene to explain non-response. Since the recognition of the importance of methylation in cancer progression, focus has now shifted to include mechanisms involving changes in methylation and their effect on drug response. Methylation could explain polygenic drug resistance, as many genes can be affected simultaneously by a small change in methylation (Glasspool, Teodoridis, and Brown 2006). For example, DNA repair enzyme 06-methyl guanine-DNA methyl transferase (MGMT) is involved in drug response in cancer, as it prevents alkylating agents from killing tumour cells (Esteller *et al.* 2000). The methylation status of the MGMT gene promoter has, therefore, become a predictor of response to alkylating agents in gliomas and colorectal cancers (Amatu *et al.* 2013; Esteller *et al.* 2000). This gene is currently the most robust methylation indicator of treatment response. Interestingly, this gene was also found to be differentially methylated between RA cases and controls in T and B lymphocytes (Glossop *et al.* 2014), making it a good possible candidate gene for therapy response studies in RA.

In addition, several methylation markers of sensitivity to chemotherapeutic drugs, most notably p73, a homologue of p53 have been identified (Shen *et al.* 2007). The hypermethylation and consequent reduced expression of this gene is associated with sensitivity to alkylating agents (Shen *et al.* 2007).
1.4.9.2. *Methotrexate and DNA methylation*

The commonly prescribed DMARD, methotrexate, depletes folate which is necessary for methylation of DNA (Chiang *et al.* 1996), indicating that treatment with methotrexate could reduce methylation of DNA. Many studies of RA have found global hypomethylation of DNA, however, one small study found that methylation levels in RA and psoriatic arthritis (PsA) reverted back to normal levels following treatment with methotrexate (Kim *et al.* 1996). As described in section 1.4.8.5, the differences in methylation between JIA cases versus healthy controls were vastly reduced following stratification for methotrexate treatment, providing further evidence that methotrexate alters DNA methylation (Ellis *et al.* 2012). Studies in cancer have also supported an association between methotrexate response and methylation levels in the *SLC19A1* gene (Worm *et al.* 2001; Yang *et al.* 2008). While this link could prove to be interesting in relation to RA treatment with methotrexate, it is important to note that RA patients receive methotrexate treatment at significantly lower doses than when it is used in cancer treatment. RA patients treated with methotrexate have also been found to have an increased risk of melanoma, lymphoma and lung cancer (Buchbinder *et al.* 2008) which could be due to dysregulation of methylation, though this has not yet been investigated.
1.5. **Summary**

RA is a complex disorder, and there has been considerable progress in recent years in identifying genes underlying the disease, however these explain only a small proportion of the heritability. Epigenetics, specifically DNA methylation, provides an alternative explanation for missing heritability. Several studies in other disease areas suggest DNA methylation could be important in disease susceptibility or in predicting treatment response but well-powered studies in RA are lacking. In methylation studies of RA the ideal study design would be longitudinal to establish temporal origins of disease and treatment response, however in studies of susceptibility, this is challenging. An alternative approach, which allows investigation into disease associated DNA methylation without confounding by genetic sequence is to investigate disease-discordant MZ twins. For treatment response, longitudinal studies are more feasible. Given health economics, prediction of response to biologic anti-TNF therapy in RA is of great importance. The development of array-based approaches have made investigation of genome-wide DNA methylation feasible, however studies employing such techniques are still limited by cost considerations. Therefore, studies investigating DNA methylation in treatment response need to be carefully designed to maximise the power to detect methylation differences, for example investigating extreme phenotypes of response. The increase in studies investigating individual cell subsets in disease has given rise to more cohorts cryopreserving cells for future separation and methylation analysis. Given the impact of environmental factors on methylation, investigation into the effects of cryopreservation of samples on DNA methylation is required to ensure the integrity of future methylation investigations.
1.6. **Aims**

The first aim of my PhD was to investigate whether epigenetic factors play a role in the response to anti-TNF biologic drug therapy in patients with RA. By comparing the methylation status of baseline DNA from patients being treated with two different anti-TNF therapies; etanercept or adalimumab who are considered good responders or non-responders, I aimed to determine whether there is an association between the epigenetic fingerprint and drug response. Patients were selected from the Biologics in RA Genetics and Genomics Study Syndicate (BRAGGSS) cohort.

In order to understand the role of epigenetics in drug response in patients with RA, it is also important to understand the underlying role of epigenetic changes in RA pathogenesis. My second aim was to investigate DNA methylation differences in RA by performing an epigenome-wide association study in RA-discordant MZ twins. Disease discordant MZ twins are an ideal study population for investigating methylation differences, as confounders which may affect epigenetic studies are controlled for, including age, gender, and genetic differences.

Many epigenetic studies are hindered by the availability of appropriate sample collections, as many existing collections of DNA are in whole blood. Epigenetics can change between cell types, making collections of cryopreserved cells from which individual cell types can later be isolated increasingly common. The third and final aim of my PhD was to investigate whether cryopreservation alters the DNA methylation profiles of cells, to determine the appropriateness of this technique for future studies.
Overall objectives:

- Patients who show an extremely good or poor response to etanercept (n=72) or adalimumab (n=72) will be selected from the BRAGGSS study. After performing bisulphite conversion on the DNA, an EWAS using the Infinium HumanMethylation450 BeadChip will be performed.
- Results from these two EWASs will be validated using pyrosequencing in a subset of samples.
- In collaboration with TwinsUK, an EWAS of RA discordant MZ twins (79 twin pairs) will be performed using whole blood or PBMC DNA.
- CD3+CD4+ T-cells extracted from fresh and cryopreserved PBMCs from 5 healthy individuals will be analysed in an EWAS performed to identify DNA methylation differences caused by cryopreservation.
2. Methods
2.1. **Sub-study sample selection**

2.1.1. **Etanercept and adalimumab response study**

Patients were selected from the Biologics in Rheumatoid Arthritis Genetics and Genomics Study Syndicate (BRAGGSS) cohort. This cohort is composed of patients with RA who were recruited at the time they were about to commence treatment with a biologic drug. All patients fulfilled the 1987 ACR criteria for classification of RA (Aletaha *et al.* 2010), at inclusion, and were Caucasian and at least 18 years of age. DNA, RNA and serum were collected prior to commencement of biologic therapy (baseline) and at 3, 6 (not RNA) and 12 (not RNA) months thereafter. Blood samples for serum, DNA and RNA extraction were received at the Arthritis Research UK (ARUK) Centre for Genetics and Genomics and entered onto a laboratory information management system (LIMS) and stored at -80°C. Psychological, demographic and clinical data, including the DAS28 (Fransen and van Riel 2005), are collected at each follow-up. Patients were selected for this study if they were treated with either etanercept or adalimumab and subsequently were assessed to have either responded well or not responded following 3 months of treatment (see section 2.1.2 for details). Where possible, patients were matched for baseline characteristics such as age, treatment with concurrent DMARDs and smoking status. Etanercept and adalimumab were selected as therapies for investigation as they are the two most widely used anti-TNF biologic therapies in the treatment of RA. All data from the BRAGGSS participants was collected following informed consent and with ethical approval (COREC 04/Q1403/37).
2.1.2. **Measuring treatment response**

Treatment response was calculated using the DAS28 score, using the following equation:

\[
\text{DAS28-CRP(4)} = 0.56\times\sqrt{\text{TJC28}} + 0.28\times\sqrt{\text{SJC28}} + 0.36\times\ln(\text{CRP}+1) + 0.014\times\text{GH} + 0.96
\]

This is a validated DAS28 formula which incorporates four variables: 28 joint tender joint count (TJC28), 28 joint swollen joint count (SJC28), C-reactive protein (CRP) and the patient's global assessment of disease activity on a 100mm VAS (GH).

Where possible, CRP was measured in baseline and three month serum samples by UK Biocentre (Cheshire, UK), otherwise CRP measurements were extracted from the patients' clinical records. Following 3 months of follow-up EULAR response criteria (Fransen and van Riel 2005) was applied to the data. Response was defined as either an endpoint DAS28<2.6 at 3 months for the good responder group (i.e. clinical remission), or an improvement of <0.6 or between 0.6-1.2 with an endpoint DAS28 of >5.1 for non-responders.

2.1.3. **Rheumatoid arthritis discordant monozygotic twins study**

Patients were selected from the Rheumatoid Arthritis Twin Study (RATS) based at the University of Manchester (Silman et al. 1993) and the TwinsUK collection based in the Department of Twin Research in King’s College London. The RA status of patients was assessed by a physician at time of sample collection. MZ twin pairs in which one twin was classified as having RA, while the other was classified as not having RA, were included in this study. Twin pairs were selected where a good quality whole blood DNA sample was available for both twins, or a cryopreserved PBMC sample was available for both twins. Ethical approval was in place (MREC99/8/84) and all samples were collected with informed consent.
2.1.4. Investigation into the effect of cryopreservation on methylation in CD3+CD4+ T-cells

Healthy volunteers were recruited from the National Repository of Healthy Volunteers (NRHV) in Manchester. The NRHV is the healthy volunteer arm of the National Repository study which is a national study that collects blood from patients with RA, unaffected family members and healthy control individuals. DNA, RNA and serum were isolated from these samples. Two fresh 10ml blood samples were collected from each participant using ethylenediaminetetraacetic acid (EDTA) vacutainer collection tubes (BD Biosciences, England) following informed consent and with ethical approval (MREC 99/8/84).
2.2. Techniques

2.2.1. Isolation of cryopreserved and fresh CD3+CD4+ T-cells

2.2.1.1. Isolation of PBMCs

Peripheral blood mononuclear cells (PBMCs) were isolated using density gradient centrifugation. This process exploits the buoyant density differences between cells to allow isolation of lymphocytes and PBMCs from whole blood. Leucosep tubes (Greiner bio-one, UK) were filled with 15ml of separation medium, Ficoll-Paque Plus (Amersham Biosciences, UK) and centrifuged for 30 seconds at 1000xg. Each 10ml blood sample was diluted with double the quantity of phosphate buffered saline (PBS) which was then poured into the leucosep tubes (Figure 2.1A). Each tube was centrifuged for 10 minutes at 1000xg in a swinging bucket rotor with the brakes switched off to prevent disruption of the PBMC layer. During this centrifugation step, the PBMCs form an enriched cell fraction which rises to the top of the separation medium due to the low density of the cells, meanwhile the unwanted erythrocytes and granulocytes form a pellet at the bottom of the tube, separated by a porous barrier (Figure 2.1B). The plasma layer was discarded leaving 10mm above the enriched cell fraction. The enriched cell fraction was then collected and washed three times using 10ml PBS followed by centrifugation for 10 minutes at 250xg.
Figure 2.1: Prepared leucosep tubes before and after centrifugation during PBMC isolation. A) Indicates the porous barrier separating the separation medium from the blood sample diluted with PBS. B) Indicates the layers which can be seen following centrifugation. The erythrocytes and granulocytes form a pellet at the base of the leucosep tubes while the enriched cell fraction lies at the top of the separation medium.
2.2.1.2. **Cryopreservation of PBMCs**

PBMC samples to be cryopreserved were resuspended using 1.8ml Recovery™ cell freezing medium (Gibco, UK) per 10ml blood sample (approximately 1x10^7 cells). This cell freezing medium contains dimethyl sulfoxide (DMSO) which reduces ice formation in the cells during freezing. Cells were then placed on ice before being transferred to -80°C in a freezing container, Mr. Frosty (Nalgene, MA, USA), containing 100% isopropyl alcohol. This allows samples to be cooled at a rate of -1°C per minute. After one week the cell aliquots were transferred to storage at between -135°C and -190°C in the vapour phase of a liquid nitrogen freezer.

2.2.1.3. **Thawing of PBMCs**

Cryopreserved PBMC samples were removed from liquid nitrogen storage and immediately thawed in a waterbath at 37°C for 5 minutes. The thawed PBMCs were transferred to pre-warmed PBS at 37°C and swirled gently to mix. The samples were centrifuged at 1600 revolutions per minute (rpm) for 7 minutes at 4°C to pellet cells and allow removal of DMSO and dead cells by discarding the supernatant.
Pellets of PBMCs were resuspended in 100µl PBS then 20µl Human Fc Receptor Binding Inhibitor added to each sample before being incubated on ice for 20 minutes. eFluor NC flow cytometry staining buffer (30µl) and 5µl each of Anti-Human CD4-eFluor450 conjugate (Affymetrix, CA, USA) and Anti-Human CD3-PE conjugate (Affymetrix, CA, USA) were added to each sample before being incubated on ice for 30 minutes. Finally 900µl of 4-2-hydroxyethyl-1-piperadineethanesulfonic acid (HEPES) buffered media with a final concentration of 25mM was added to each sample before filtering through a 50-micron filter within a sterile Cup-Type Filcon (#340630 BD Biosciences, England). Following sample preparation, CD3+CD4+ T cells were isolated by FACS performed at the Faculty of Life Sciences core facility in the University of Manchester by Mr Michael Jackson, using a FACSARia™ (BD Biosciences, England).
2.2.2. DNA extraction

2.2.2.1. Manual method of whole blood DNA extraction

Whole blood DNA for the anti-TNF response studies and the discordant RA twin study were extracted from 10ml blood and normalised by Arthritis Research UK Centre for Genetics and Genomics lab technicians.

Red blood cells were lysed by adding EDTA blood samples to double the volume of lysis buffer (containing 0.15M NH₄Cl, 0.0092M KHCO₃, 0.002M EDTA and dH₂O) in a fresh 50ml falcon tube (Fisher Scientific, UK). After briefly vortexing, samples were mixed on a roller for 10 minutes then centrifuged at 3040xg for 10 minutes (Rotina 420, Hettich). The supernatant was discarded and the pellet resuspended in 20ml 2x lysis buffer. Samples were vortexed then mixed on a roller for 10 minutes. Samples were centrifuged at 3040xg for 10 minutes then the supernatant was discarded. DNA was then extracted from the white cell pellet using Phenol:Chloroform:Isoamylalcohol. Salt/EDTA buffer (3ml), 300µl 10% sodium dodecyl sulphate (SDS) and 50µl 10mg/ml Proteinase K (Sigma-Aldrich, UK) was added to each white cell pellet, then each sample was vortexed and incubated overnight to allow digestion of the cellular proteins. After overnight incubation, samples were vortexed and 3ml phenol/chloroform/isoamyl alcohol (25:24:1) was added to each sample. Samples were vortexed until a homogenous white emulsion was produced, then left to stand for one minute. Samples were centrifuged at 3040xg for 20 minutes at room temperature. The aqueous upper layer was collected with a pastette then a further 3ml phenol/chloroform/isoamyl alcohol (25:24:1) was added to each sample. After centrifugation at 3040xg for 10 minutes, the aqueous upper layer was transferred to 6ml ice-cold absolute ethanol. Samples were left at -20°C overnight to allow precipitation of the DNA. Following overnight precipitation, the samples were inverted 50 times then centrifuged at 3040xg for 10 minutes to pellet the DNA. The pellets were washed with 2ml 70% ethanol, the ethanol
discarded and the pellets allowed to air-dry for one hour until no ethanol remained. Pellets were resuspended in 300µl T_{10}E_{0.1} (10mM Tris-HCl, pH 8.0, 0.1mM EDTA, pH 8.0). Samples were mixed overnight at 300rpm on a shaker (IKA KS 260).

2.2.2.2. Phase-lock gel columns method

I extracted DNA for use in the investigation of the effects of cryopreservation on DNA methylation, and from PBMC samples in the RA discordant twin study.

DNA was extracted from CD3+CD4+ T-cells using 2ml light phase-lock gel (PLG) columns (5prime, Germany). Samples were centrifuged for 10 minutes at 4000rpm to pellet the cells then supernatant was removed. Samples were resuspended in 500µl salt EDTA buffer, 60µl 10% SDS and 20µl proteinase K then incubated at 37°C overnight. Prior to use, the PLG columns were centrifuged at 14000xg for 30 seconds in a microcentrifuge to pellet the gel. The aqueous sample and an equal quantity of organic extraction solvent (phenol/chloroform/isoamyl alcohol 25:24:1) was added to the PLG tube and the organic and aqueous phases thoroughly mixed to a homogenous suspension by inverting the tube. The tubes were centrifuged for 5 minutes at 14000xg to separate the phases then the aqueous upper layer containing the nucleic acid was transferred to a new 2ml cryovial and double the volume of ice cold absolute ethanol added. The samples were incubated on ice for 45 minutes to allow the DNA to precipitate out of solution before centrifuging the sample at 14000xg for 15 minutes to pellet the DNA. The supernatant was discarded and samples were washed using 70% ethanol followed by centrifugation at 14000xg for 15 minutes. The supernatant was discarded then the pellet was left to air dry until no visible ethanol remained. Pellets were resuspended in 50µl T_{10}E_{0.1} buffer with shaking overnight on the KS 260 shaker (Ika, NC, USA).
2.2.3. **Quality control of DNA**

2.2.3.1. *Gel electrophoresis*

To visualise sample degradation and allow selection of the highest quality samples for further experiments, DNA was run on 1% agarose gels containing 2µl ethidium bromide (Life technologies, UK). For each sample, 2µl genomic DNA was added to 2µl loading buffer and electrophoresed on a Hybaid Electro-4 Gel System (Fisher Scientific, UK) at 120V for 50 minutes. The bands were visualised using a transilluminator (UVP, CA, USA) and were considered good quality if a strong band was clear at approximately 10-20 kb (example of good and poor DNA quality shown in Figure 2.2).

![Figure 2.2: Example of good and poor quality DNA sample, assessed using gel electrophoresis.](image)

Figure 2.2: Example of good and poor quality DNA sample, assessed using gel electrophoresis.
2.2.3.2. **Nanodrop**

To measure the concentration of DNA, samples were analysed using a NanoDrop1000 Spectrophotometer V3.2.1 (Thermo Scientific, Wilmington, DE). This instrument measures the absorbance of light at wavelengths of 260nm and 280nm. By creating a ratio of these two measurements, the concentration and purity of DNA in the sample is estimated. A ratio of approximately 1.8 is expected for pure DNA. An initial blank measurement of 1.2µl of purified water followed by 1.2µl of TE buffer was used to calibrate the machine. After cleaning the upper and lower pedestals using lint free tissue, 1.2µl of each sample was loaded onto the lower measurement pedestal, and the sampling arm was closed allowing the spectral measurement to be made.

2.2.3.3. **Dropsense**

To obtain a more accurate estimation of sample composition and presence of contaminants, DNA samples were also analysed using the Dropsense96 UV/VIS reader (Trinean, Belgium). This method quantifies nucleic acid using spectral analysis (230-750nm). Samples were then analysed using cDrop software (Trinean, Belgium) to measure the concentration of samples and the degree of contamination from RNA and phenol.
2.2.4. **Bisulphite conversion**

2.2.4.1. **Principles**

Bisulphite conversion is a process that allows the differentiation of methylated and unmethylated cytosine residues by converting the epigenetic signature of DNA methylation into a genetic signature. The process relies on the different rates at which methylated and unmethylated cytosines are deaminated by sodium bisulphite. Unmethylated cytosines are deaminated by sodium bisulphite and converted to uracil residues much more readily than methylated cytosines (Figure 2.3). During PCR amplification, uracil residues are replaced by thymine residues (Figure 2.4). Subsequent analysis of the DNA sequence allows determination of whether cytosine residues were methylated or not (Hayatsu 2008).

![Figure 2.3](Image)

Figure 2.3: Mode of action of sodium bisulphite (HSO3-) in the bisulphite conversion of DNA. Unmethylated cytosine residues (left) are deaminated by sodium bisulphite to uracil, while methylated cytosine residues (right) are not affected by sodium bisulphite and so remain as cytosine residues.
Figure 2.4: Example of translation of epigenetic signature into a genetic signature during bisulphite conversion of DNA and subsequent PCR. Methylated cytosine is shown to remain as a cytosine residue during bisulphite conversion and PCR however unmethylated cytosine in a CpG context and non-CpG context is deaminated to uracil residue during bisulphite conversion which is then translated into thymine during PCR.
2.2.4.2. **Method**

DNA samples were bisulphite converted using an EZ DNA Methylation Kit from Zymo Research, CA, USA. The amended Zymo protocol (Zymo Research 2012a) recommended by Illumina for use with the Infinium HumanMethylation450 BeadChip was used.

In total 5µl of M-dilution buffer was added to 500ng DNA from each sample, after which the sample volume was made up to 50µl with milliQ water (Merck KGaA, Darmstadt, Germany). After incubating the sample for 15 minutes at 37°C on a heat block (HybEx Heating System, Illumina, USA), 100µl of CT Conversion Reagent (Zymo, CA, USA) was added to each sample. The samples were then incubated in a thermocycler (GeneAmp® PCR system 9700 from Applied Biosystems) for 16 cycles of 95°C for 30 seconds followed by 50°C for 60 minutes. Following 10 minute incubation at 4°C, the samples were loaded into Zymo-Spin™ IC Columns containing 400µl of M-Binding Buffer. This allows the DNA to bind to the pad at the base of the spin column. Samples were centrifuged for 30 seconds at 10,000xg in a microcentrifuge, and the flow-through discarded. Samples were washed twice with 200µl M-Wash Buffer. After transferring the column to a 1.5ml microcentrifuge tube, the sample was eluted in 12µl of M-Elution Buffer and centrifuged for 30 seconds at 10,000xg. Samples were stored at -20°C.
2.2.5. **Epigenome-wide association studies**

2.2.5.1. **Principles**

Whole genome methylation was measured using the Illumina HumanMethylation450 BeadChip. This technology measures methylation levels at 485,577 CpG sites and relies on bisulphite conversion of the DNA samples (section 2.2.4). The CpG sites on the array cover 99% of RefSeq genes and 96% of CpG islands, with an average of 17 CpG sites per gene region (Illumina 2014). The array employs Infinium I and II chemistry to measure methylation status of targeted CpG loci. Infinium I chemistry employs two probes for each CpG site, one designed to target the methylated locus (M bead type) and one designed to target the unmethylated locus (U bead type) (Figure 2.5). Infinium II chemistry employs a single probe for each targeted CpG, which differentiates between the methylated and unmethylated locus depending on the colour of the ddNTP incorporated during single base extension (Figure 2.5). By measuring the fluorescent signals emitted by each probe, the methylation level at each locus can be calculated.
Figure 2.5: Summary of Infinium I and Infinium II chemistries present on the Illumina HumanMethylation450 BeadChip.
Infinium I chemistry involves two probe types for each methylation state while Infinium II involves just one probe type for each methylation state. Unmethylated Infinium I probes (U) at an unmethylated locus will fully bind to the DNA allowing single base extension, U probes will not bind to a methylated locus so single base extension will not occur. Infinium II probes will bind to DNA of either methylation state, however methylated or unmethylated DNA will produce a different coloured signal (adapted from Illumina 2014).
Each BeadChip can measure DNA methylation levels of 12 samples in a 2x6 matrix, with each sample occupying one array (Figure 2.6). For each sample there are 600 negative control probes as well as bisulphite conversion controls and sample-independent controls for staining, extension, hybridisation and target removal. Samples were prepared using the manufacturers recommended manual protocol (Illumina 2011). The individual steps for preparing samples are described in detail below in sections 2.2.5.2-2.2.5.11.

Figure 2.6: Photo of the Illumina HumanMethylation450 BeadChip. The HumanMethylation450 BeadChip allows measurement of DNA methylation in 12 samples arranged in a 6x2 matrix (Illumina 2014).
2.2.5.2. **Denaturation, neutralization and amplification of samples**

For each sample, 4µl of bisulphite converted DNA (approximately 200ng) was transferred to a well of a 0.8ml microtiter plate (#AB-0765, Thermo Scientific Abgene, UK), to which 20µl multi-sample amplification 1 mix (MA1, Illumina) was added. To each well, 4µl 0.1M sodium hydroxide (NaOH) was added before the plate was sealed with an adhesive seal (Thermo Scientific, UK) and vortexed at 1600rpm for one minute (High-speed microplate shaker, Illumina). Following pulse centrifugation to 280xg, the samples were incubated for 10 minutes at room temperature. Random primer mix (68µl) (Illumina) was added to each sample followed by 75µl multi-sample amplification master mix (MSM, Illumina) then the plate re-sealed. The samples were vortexed at 1600rpm for 1 minute, then pulse centrifuged to 280xg. Samples were incubated at 37°C for 20-24 hours in the Illumina Hybridisation oven (#SE-901-1001, Illumina).

2.2.5.3. **Fragmentation of amplified samples**

The plate containing the samples was pulse centrifuged to 280xg before 50µl fragmentation solution (FMS, Illumina) was added to each sample-containing well. The plate was sealed then vortexed at 1600rpm for 1 minute before being pulse centrifuged. Samples were incubated at 37°C for 1 hour using the HybEx Heating System (Illumina).
2.2.5.4. **Precipitation of samples**

Precipitation solution (100µl) (PM1) was added to each sample, then the plate was sealed and vortexed at 1600rpm for 1 minute. Samples were incubated at 37°C for 5 minutes then pulse centrifuged to 280xg before 300µl of 100% 2-propanol was added. The plate was sealed with a foil seal (Thermo Scientific, UK) then inverted 10 times to mix the contents. The samples were incubated at 4°C for 30 minutes then centrifuged at 3,000xg at 4°C for 20 minutes. Over absorbant paper roll, the supernatant was removed from the sample pellet by quickly inverting the plate then firmly tapping the plate several times; avoiding areas where supernatant had already been absorbed. The plate was then left inverted on a rack for one hour at room temperature to air dry the pellets.

2.2.5.5. **Resuspension of samples**

To each sample pellet 46µl resuspension, hybridisation and wash solution (RA1, Illumina) was added before the plate was sealed with a foil seal. The samples were incubated for 1 hour at 48°C, then vortexed at 1800rpm for 1 minute followed by pulse centrifugation to 280xg.

2.2.5.6. **Hybridization of samples to BeadChips**

The samples were denatured by heating the plate to 95°C for 20 minutes. In preparation for hybridisation, 400µl humidifying buffer (PB2, Illumina) was dispensed into the reservoirs of the BeadChip Hybridisation Chamber (Illumina), and covered with the chamber lid to prevent evaporation. The plate containing the samples was then left at room temperature for 30 minutes, then pulse centrifuged to 280xg. Each individual BeadChip was placed into a Hyb Chamber insert (Illumina) then 10µl of each DNA sample was dispensed onto the appropriate BeadChip
section. The Hyb Chamber inserts containing the BeadChips were then loaded into the Hyb Chamber and the lid clamped down. Each Hyb Chamber accommodates four BeadChips. The BeadChips were incubated for 16 hours at 48°C in the Illumina Hybridisation Oven.

2.2.5.7. *Washing the BeadChips*

The BeadChips were removed from the Hyb Chambers, and the cover seal carefully removed before being submerged in BeadChip preparation reagent (PB1, Illumina) using the wash rack (Illumina). The wash rack was gently moved up and down in the PB1 for 1 minute, breaking the surface of the PB1. The wash rack was then transferred to a wash dish containing clean PB1, and the rack was again moved up and down breaking the surface, for 1 minute. Each BeadChip was then transferred to a black frame in the Multi-sample BeadChip Alignment Fixture (Illumina) containing 150ml PB1. A clear spacer was placed on top of each BeadChip, then BeadChips were aligned using the Alignment Bar. A clean glass back plate was placed on top of the spacer of each BeadChip and was clamped in place creating a Flow-Through Chamber for each BeadChip (Figure 2.7). The ends of the clear plastic spacers were then trimmed off using scissors.
Figure 2.7: Flow-through chamber assembly of BeadChip used during single-base extension and staining steps of BeadChip preparation. Each BeadChip is placed in a black frame, then a spacer is placed on top followed by a glass back plate, then the setup is secured with clamps. The glass back plate has a slanted gap at the top, which when in contact with the BeadChip creates a reservoir for reagents to be added to during single-base extension and staining. The spacers provide a slight gap in between the BeadChip and glass back plate allowing reagents to run over the arrays.
### Single base extension

Each Flow-Through Chamber assembly was placed into a Chamber Rack (Illumina) and incubated at 44°C, before the reagents shown in Table 2.1 were dispensed to each BeadChip using a 1000µl pipette into the reservoir of each Flow-Through Chamber in the order shown in Table 2.1.

<table>
<thead>
<tr>
<th>Reagent</th>
<th>Quantity</th>
<th>Incubation</th>
<th>Repeats</th>
</tr>
</thead>
<tbody>
<tr>
<td>RA1</td>
<td>150µl</td>
<td>30 seconds</td>
<td>5</td>
</tr>
<tr>
<td>XC1</td>
<td>450µl</td>
<td>10 minutes</td>
<td>None</td>
</tr>
<tr>
<td>XC2</td>
<td>450µl</td>
<td>10 minutes</td>
<td>None</td>
</tr>
<tr>
<td>TEM</td>
<td>200µl</td>
<td>15 minutes</td>
<td>None</td>
</tr>
<tr>
<td>95% formaldehyde/1mM EDTA</td>
<td>450µl</td>
<td>1 minute</td>
<td>1</td>
</tr>
</tbody>
</table>

Table 2.1: Order, name and quantity of reagents added to Flow-Through chamber in initial stages of single-base extension during BeadChip preparation. RA1: Resuspension, hybridization and wash solution, XC1: Xstain BeadChip solution 1, XC2: Xstain BeadChip solution 2, TEM: two-colour extension master mix.

Each Flow-Through Chamber was then incubated for 5 minutes before the Chamber Rack temperature was ramped down to 34°C. Xstain BeadChip solution 3 (450µl) (XC3, Illumina) was added then incubated for 1 minute (this was repeated once).
2.2.5.9. **Staining of BeadChips**

When the Chamber Rack reached 34°C, the BeadChips were stained by adding the reagents shown in Table 2.2.

<table>
<thead>
<tr>
<th>Reagent</th>
<th>Quantity</th>
<th>Incubation</th>
<th>Repeats</th>
</tr>
</thead>
<tbody>
<tr>
<td>STM</td>
<td>250 µl</td>
<td>10 minutes</td>
<td>None</td>
</tr>
<tr>
<td>XC3</td>
<td>450 µl</td>
<td>1 minute</td>
<td>1</td>
</tr>
<tr>
<td><strong>Wait 5 minutes</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ATM</td>
<td>250 µl</td>
<td>10 minutes</td>
<td>None</td>
</tr>
<tr>
<td>XC3</td>
<td>450 µl</td>
<td>1 minute</td>
<td>1</td>
</tr>
<tr>
<td><strong>Wait 5 minutes</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>STM</td>
<td>250 µl</td>
<td>10 minutes</td>
<td>None</td>
</tr>
<tr>
<td>XC3</td>
<td>450 µl</td>
<td>1 minute</td>
<td>1</td>
</tr>
<tr>
<td><strong>Wait 5 minutes</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ATM</td>
<td>250 µl</td>
<td>10 minutes</td>
<td>None</td>
</tr>
<tr>
<td>XC3</td>
<td>450 µl</td>
<td>1 minute</td>
<td>1</td>
</tr>
<tr>
<td><strong>Wait 5 minutes</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>STM</td>
<td>250 µl</td>
<td>10 minutes</td>
<td>None</td>
</tr>
<tr>
<td>XC3</td>
<td>450 µl</td>
<td>1 minute</td>
<td>1</td>
</tr>
<tr>
<td><strong>Wait 5 minutes</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 2.2: Order of addition, incubation periods and repeats for each reagent in staining protocol during preparation of BeadChips. STM: superior two-colour master mix, XC3: XStain BeadChip solution 3, ATM: anti-stain two-colour master mix.

The Flow-Through Chambers were immediately removed from the Chamber Rack and placed horizontally at room temperature.
2.2.5.10. *Washing and coating the BeadChips*

The metal clamps were removed from the Flow-Through Chamber using a dismantling tool then the glass back plate and spacer were removed. The BeadChips were placed into the staining rack and submerged in a wash dish containing 310ml PB1. The staining rack was moved up and down 10 times, breaking the surface of the PB1 each time. The BeadChips were left to soak in PB1 for 5 minutes before being transferred to a wash dish containing 310ml XStain BeadChip solution 4 (XC4, Illumina) and moved up and down 10 times, then left to soak for 5 minutes. The staining rack was removed from the wash dish and placed with the BeadChips facing upwards, which were subsequently removed using locking tweezers. BeadChips were left to dry on a tube rack in a vacuum desiccator for 50 minutes at 675 mmHg. The underside of the BeadChips was cleaned using ethanol wipes.

2.2.5.11. *Imaging the BeadChips*

The BeadChips were loaded into the BeadChip Carrier (Illumina) then settled into position by pressing the Lift button at the side of the carrier. Using the iScan Control Software (ICS), the tray was opened and the carrier was lined up with the adapter plate of the iScan Reader Tray. The iScan Reader Tray was closed, and after the barcodes were scanned by the iScan, ICS setup was completed. Once the ICS successfully performed tilt and align functions, the BeadChips were scanned.
2.2.6. **Pyrosequencing**

2.2.6.1. **Principle of pyrosequencing**

Pyrosequencing is a method of generating quantitative methylation measurements of small sequences of DNA using a sequencing-by-synthesis method. Bisulfite converted DNA is amplified by PCR in the region of interest, using a biotin-labelled primer set. The PCR amplicon is denatured and made single stranded (the biotin-labelled strand is selectively retained through the use of streptavidin coated sepharose beads) before being annealed to the sequencing primer (Figure 2.8A). Deoxyribo-nucleotide triphosphate (dNTP) residues are incorporated if they complement the single stranded template, in a reaction catalysed by DNA polymerase. Each time a nucleotide is incorporated, equal amounts of pyrophosphate (PPI) are produced (Figure 2.8B). This PPI is converted to ATP by ATP sulfurylase in the presence of adenosine 5' phosphosulfate (APS) (Figure 2.8C). The ATP produced is needed as a coenzyme for the conversion of luciferin to oxyluciferin by the enzyme luciferase. This reaction produces light which is proportional to the amount of ATP and hence the amount of nucleotide incorporated (Figure 2.8D). The light emitted is measured by a charge coupled device (CCD) which translates the signal to produce a Pyrogram, the raw data output file of pyrosequencing (Figure 2.8E). In the resultant pyrogram, each peak is proportional to the amount of nucleotide incorporated (Figure 2.8F).
Figure 2.8: Principles of pyrosequencing technology.
A) After PCR amplification of the region of interest, a single stranded product is produced and annealed to the sequencing primer. B) This single stranded product is extended by incorporation of deoxribonucleotide triphosphate residues (dNTPs), catalysed by DNA polymerase. Each nucleotide incorporated produces equimolar pyrophosphate (PPI). C) PPI is converted to ATP by ATP sulfurylase in the presence of adenosine 5' phosphosulfate (APS). D) Luciferin is converted to oxyluciferin by luciferase in an ATP dependent reaction. E) The production of oxyluciferin generates visible light which is detected using a charge coupled device (CCD). F) The program is produced from this light measurement, with each peak being proportional to the number of nucleotides incorporated.
2.2.6.2. *Pyrosequencing assay design*

Assays used for the pyrosequencing validation of 450k data were designed using the Methylation Analysis Module of the PyroMark Assay Design Software 2.0 (Qiagen, UK). Sequences flanking each CpG targeted for validation were extracted from the HumanMethylation450 manifest file. Using the Basic Local Alignment Search Tool-Like Alignment Tool (BLAT, UCSC), this sequence was extended by 150bp upstream and downstream, giving a larger sequence of 422bp to design assays from. This sequence was input into the ‘Original Sequence Editor’ tab of the PyroMark Assay Design Software. The ‘target region’ was selected to include the CpG of interest and the automatic assay design was used to generate primer sets which include a forward and reverse primer to be used for amplification of DNA, and a sequencing primer. A ‘Primer Set Score’ was automatically generated for each primer set, and the primer set with the highest score was selected for each assay. A score of 88 or more is considered a high quality assay. Primer sets are given a lower score if the software detects a problem such as possible hairpin loop formation or sub-optimal amplicon length.
2.2.6.3. **PCR to create biotin labelled amplicons for pyrosequencing**

Biotin-labelled amplicons were generated in a PCR reaction using the custom PCR primer set generated during assay design. Each reaction contained 12.5µl PyroMark PCR Mastermix 2x (Qiagen), 2.5µl CoralLoad Concentrate 10x (Qiagen), 2.5µl PCR primer set 10x (custom assay from QIAGEN) and 10-20ng DNA, made up to 25µl with RNase-free water. The PCR mastermix contains HotStarTaq DNA polymerase, and an optimized PCR buffer containing dNTPs and 3mM MgCl₂. The CoralLoad Concentrate improves specificity of PCR and acts as a loading buffer during agarose gel analysis of samples. Using a GeneAmp9700 Thermocycler (Applied Biosystems), PCR was performed using the following cycles: 95°C for 15 minutes, [94°C for 30 seconds, 56°C for 30 seconds, 72°C for 30 seconds] x45 cycles, followed by 10 minutes at 72°C.

2.2.6.4. **Quality check of PCR amplicon size**

To confirm the presence of a correctly sized amplicon, 5µl of the PCR amplified DNA sample was electrophoresed on a 2% agarose gel containing 4µl ethidium bromide for 1 hour at 120V. Loading buffer was not necessary as CoralLoad Concentrate acts as loading buffer. Gels were then visualised using a transilluminator.

2.2.6.5. **Creation of a single stranded amplicon using vacuum workstation**

The ‘vacuum workstation function test’ was carried out to ensure all probes of the handset were functioning. This involves flushing the vacuum handset with 70ml milliQ water, followed by 100µl milliQ water from each well of a 24 well shallow plate (Qiagen).
After gently inverting the streptavidin coated sepharose beads for two minutes, 20µl of the PCR product was added to 2µl sepharose beads, 18µl milliQ water and 40µl binding buffer in a mastermix which was then distributed among the samples in a 96-well PCR plate. Once all components were combined they were mixed with shaking at 600rpm for 10 minutes using a High-speed microplate shaker (Illumina). While the PCR product and bead mix was shaking, the vacuum workstation was set up as shown in Figure 2.9, using (1) 50ml 70% ethanol, (2) 40ml PyroMark Denaturation Solution (Qiagen), (3) 50ml 1x PyroMark Wash Buffer (Qiagen), (4) 50ml and (5) 70ml milliQ water.

Figure 2.9: Setup of the vacuum workstation used to create single stranded amplicons for use in pyrosequencing. The arrows indicate the order in which the vacuum handset should be applied to each reagent, starting with the biotin labelled PCR product.
The sequencing primer was prepared by adding 2.5µl sequencing primer (x10) and 22.5µl annealing buffer to each well of a 24 shallow well plate (Qiagen). The vacuum was applied to the vacuum tool and after flushing the vacuum handset with 70ml of milliQ water, it was lowered into the biotin labelled PCR product for 15 seconds. The handset was flushed with 70% ethanol for 5 seconds, followed by denaturation solution for 5 seconds, and wash buffer for 10 seconds. After draining the tool at a 90˚ angle for 5 seconds, the vacuum was switched off and the beads attached to the filter probes were released into the sequencing primer solution. To allow the biotin-labelled DNA strand to anneal to the primers, the solution was heated to 80˚C for 2 minutes on a pre-heated thermocycler block.
2.2.6.6. **Pyrosequencing using PyroMark Q24**

The PyroMark Q24 Cartridge was prepared using the PyroMark Gold Q24 Reagents (Qiagen) in the order shown in Figure 2.10. Enzyme mixture, substrate mixture and the four deoxynucleotides (dATP, dCTP, dGTP and dTTP) were added in the quantities specified for each individual reaction by the PyroMark Q24 Software.

Figure 2.10: Location of PyroMarks Gold Q24 reagents in the cartridge.
E: enzyme mixture, S: substrate mixture, A: dATP, C: dCTP, G: dGTP and T: dTTP
2.2.6.7. **Quality control of bisulphite conversion and validation of pyrosequencing technology with Sanger sequencing**

This thesis describes bisulphite conversion and pyrosequencing for methylation analysis, both of which were new techniques in our group. To test efficiency and reliability of the techniques the following quality control experiments were performed prior to processing samples for the treatment response and twin studies described.

To assess bisulphite conversion efficiency and to validate pyrosequencing with Sanger sequencing, methylated DNA standard Human HCT116 DKO (Zymo Research, Irvine, CA), non-methylated DNA standard Human HCT116 DKO (Zymo Research, Irvine, CA) and three whole blood DNA samples from the BRAGGSS cohort (see section 2.1.1) were analysed. The samples were bisulphite converted (see section 2.2.4) in duplicate to allow comparison of technical and biological replicates.

The level of DNA methylation for each sample was then analysed using both pyrosequencing and Sanger sequencing of the death-associated protein kinase 1 (DAPK1) gene. The region was recommended by the manufacturers as a positive control for checking bisulphite conversion efficiency (Zymo Research 2012b). The region selected for Sanger sequencing was too long for a single pyrosequencing assay and so was divided into two assays (referred to as DAPK1_3 and DAPK1_5).

A third assay, pre-designed by Qiagen and targeting a different region of the DAPK1 gene (referred to as DAPK1_1) was also used. In analysis of both DAPK1 regions, bisulphite conversion efficiency was measured by assessing the measurements at non-CpG cytosines, which would always be expected to convert to a thymine residue. Reliability of pyrosequencing was assessed by comparing technical replicates.
2.2.6.7.1.  **Sanger sequencing of DAPK1**

PCR primers provided in the Human Methylated & Non-methylated DNA set (Zymo Research) were used to amplify a 274bp region of *DAPK1* in preparation for Sanger sequencing (sequence shown in Table 2.3). In total, 0.5µM of the PCR primers were combined with 20ng of bisulphite converted DNA, 0.2mM of each dNTP, 2 units of ZymoTaq DNA polymerase (Zymo Research) and 1x standard reaction buffer (supplied with ZymoTaq™ DNA polymerase). The solution was made up to 20µl with DNase/RNase-Free water (Zymo Research) before the PCR was performed in a thermocycler (GeneAmp® PCR system 9700 from Applied Biosystems) following the conditions shown below (Table 2.4).

<table>
<thead>
<tr>
<th>Primer name</th>
<th>Primer sequence</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>DAPK1</em> primer I</td>
<td>ATTGGGAAGGTAAAGGYYGGAGGAAATTTGGT</td>
</tr>
<tr>
<td><em>DAPK1</em> primer II</td>
<td>CCCAAACRCAAACATCCCAAAACCAACCACATTCCTA</td>
</tr>
</tbody>
</table>

Table 2.3: Primer sequences from Human Methylated & Non-Methylated DNA set from Zymo Research, used in Sanger sequencing of bisulphite converted DNA. Sequences are displayed in the 5’ to 3’ direction.

<table>
<thead>
<tr>
<th>Temperature</th>
<th>PCR step</th>
<th>Time</th>
</tr>
</thead>
<tbody>
<tr>
<td>95°C</td>
<td>Initial denaturation</td>
<td>10 minutes</td>
</tr>
<tr>
<td>95°C</td>
<td>Denaturation</td>
<td>30 seconds</td>
</tr>
<tr>
<td>59°C</td>
<td>Annealing</td>
<td>60 seconds</td>
</tr>
<tr>
<td>72°C</td>
<td>Extension</td>
<td>60 seconds</td>
</tr>
<tr>
<td>72°C</td>
<td>Final Extension</td>
<td>7 minutes</td>
</tr>
<tr>
<td>4°C</td>
<td>Hold</td>
<td></td>
</tr>
</tbody>
</table>

Table 2.4: Thermocycler conditions used for the amplification of bisulphite converted DNA using *DAPK1* primers in preparation for Sanger sequencing.
To eliminate unincorporated primers and dNTPs, 5µl from each PCR reaction was mixed with 2µl ExoSAP-IT (Affymetrix) and incubated at 37°C for 15 minutes followed by 80˚ for 15 minutes. 10ng DNA from each sample was then added to 4pmoles of the forward and reverse sequencing primers separately (Metabion International AG), the sequences of which are shown below (Table 2.5). Two sequencing primers are used as each region under investigation undergoes forward and reverse sequencing in order to aid with sequence alignment. The reaction mix was then made up to 10µl with purified water and sent to The University of Manchester DNA Sequencing Facility for sequencing.

<table>
<thead>
<tr>
<th>Primer name</th>
<th>Primer sequence</th>
</tr>
</thead>
<tbody>
<tr>
<td>DAPK1 forward</td>
<td>ATT GGG AAG GTT AAG GYG GAG GGA AAT TTG GT</td>
</tr>
<tr>
<td>DAPK1 reverse</td>
<td>CCC CAA ACR AAA CAA TCC CCA AAA CCA CAT TCC TA</td>
</tr>
</tbody>
</table>

Table 2.5: Sequences of forward and reverse sequencing primers for the sequencing of DAPK1 using Sanger sequencing.

The results from the Sanger sequencing were analysed using FinchTV version 1.4 (PerkinElmer, Inc). Due to inconsistencies in sequencing at the start and end of the sequences, a 50bp region in the middle of the sequence was selected for analysis of bisulphite conversion efficiency. CpG sites were interrogated for suggested methylation (represented by a C residue at these sites) and non-CpG cytosines were analysed to measure bisulphite conversion efficiency. At these sites, if bisulphite conversion had been efficient, all bases would be expected to be converted to thymine residues.
2.2.6.7.2. *Pyrosequencing of DAPK1*\_3 and DAPK1\_5

The region of the *DAPK1* gene sequenced using Sanger sequencing was also analysed using pyrosequencing. Due to the length of the region, it was too long for a single pyrosequencing assay, so it was split into two pyrosequencing assays, DAPK1\_3 and DAPK1\_5, the sequences of which are shown below (Table 2.6). Pyrosequencing was performed using the protocol described in sections 2.2.6.3-2.2.6.6.

<table>
<thead>
<tr>
<th>Assay name</th>
<th>DAPK1_3</th>
<th>DAPK1_5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Forward PCR primer</td>
<td>GGGTAGGTGTATTTAGGAATTTGGT</td>
<td>GGAGGGATAGGGGGAGTTTT</td>
</tr>
<tr>
<td>Reverse PCR primer</td>
<td>TAACCACCTCTCCCCCATTCC</td>
<td>CCCCAAACCACATTCCT</td>
</tr>
<tr>
<td>Sequencing Primer</td>
<td>CCTCTCCCTCCCC</td>
<td>GGTGTGTTAGGGTAGTGT</td>
</tr>
</tbody>
</table>

Table 2.6: Primer sequences used in pyrosequencing of two regions of the DAPK1 gene, DAPK1\_3 and DAPK1\_5. All primer sequences are listed in the 5’ to 3’ direction.

2.2.6.7.3. *Pyrosequencing of DAPK1\_1*

Due to poor performance of the DAPK1\_3 and DAPK1\_5 assays, another region of the DAPK1 gene was sequenced using a pre-designed pyrosequencing assay, referred to as DAPK1\_1 (Qiagen, United Kingdom). The assay was performed on the same 5 samples (described in section 2.2.6.7) in duplicate using the protocol described in sections 2.2.6.3-2.2.6.6. As this was a pre-designed assay, the primer sequences are not available, however the sequence to be analysed contained 7 CpG sites and maps to chromosome 9: 90,113,501-90,113,542.
2.2.6.8. Validation of epigenome-wide association studies using pyrosequencing

In order to perform a technical validation of the HumanMethylation450 BeadChip methylation results, the methylation of the top 5 most differentially methylated positions from the two treatment response studies was measured using pyrosequencing, using the protocol described in sections 2.2.6.2-2.2.6.6. The assays used in these validation studies are described in sections 3.2.1.3 and 3.2.2.3.
2.3. **Statistical analysis**

2.3.1. **Epigenome-wide association studies of anti-TNF response**

2.3.1.1. **Overview**

All analysis was performed using GenomeStudio analysis software (Illumina) and using the Bioconductor v2.14 analysis packages minfi (Hansen and Aryee 2013), PCA methods (Stacklies *et al.* 2007), SVA (Leek *et al.* 2013) and CpG.assoc (Barfield, Conneely, and Kilaru 2012). A summary of the analysis pipeline is shown in Figure 2.11. Initial quality checks were performed using GenomeStudio and the minfi package to remove any poor quality samples and probes. Individual probes were removed if they contained SNPs, or if they mapped to the X or Y chromosomes. Probes were assessed for quality using detection p-values. Detection p-values were calculated for all probes by comparing the combined methylated and unmethylated signal at each position with the background signal level, as estimated using negative control positions (Hansen and Aryee 2013). Following Subset Within-Array Normalisation (SWAN), principal components analysis (PCA) was used to identify any batch effects which were subsequently adjusted for using the ComBat function within the SVA package. Differentially methylated positions were identified using a mixed effects model in the package CpG.assoc, adjusting for clinical variables that have previously been shown to correlate with treatment response (Emery and Dorner 2011; Hyrich *et al.* 2006; Kristensen *et al.* 2008) as fixed effects. In the separate analyses of etanercept and adalimumab response, methylation array was included as a random effect, while this was changed to position on array for the combined analysis of the two drugs.
Figure 2.11: Summary of methylation analysis pipeline.

- **Quality control**
  - Identify and remove poor samples

- **SWAN normalisation**

- **Pruning unwanted probes:**
  - failed detection P (>0.01)
  - gender (XY) probes
  - probes containing SNPs

- **Principal Components Analysis (PCA)**
  - Using PCA methods
  - to identify batch effects

- **ComBat**
  - to adjust batch effect

- **Identify differentially methylated positions using cg.assoc**
  - mixed effects model
2.3.1.2. **Quality control**

Initial quality checks were performed in GenomeStudio (Illumina) by visually studying the control dashboard, which summarises the performance of sample-dependent and sample-independent control probes present on the array. The sample-independent controls measure the performance of specific steps in the array preparation, including staining, extension, target removal and hybridisation. The sample-dependent controls measure the performance of sample specific characteristics, including bisulphite conversion controls (type I and type II), specificity controls (type I and type II), non-polymorphic controls and negative controls. The expected intensity of each type of control probe is shown in Table 2.7.
<table>
<thead>
<tr>
<th>Purpose</th>
<th>Name</th>
<th>Number on array</th>
<th>Expected intensity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Staining</td>
<td>DNP (High)</td>
<td>1</td>
<td>High</td>
</tr>
<tr>
<td></td>
<td>DNP (Background)</td>
<td>1</td>
<td>Background</td>
</tr>
<tr>
<td></td>
<td>Biotin (Medium)</td>
<td>1</td>
<td>High</td>
</tr>
<tr>
<td></td>
<td>Biotin (Background)</td>
<td>1</td>
<td>Background</td>
</tr>
<tr>
<td>Extension</td>
<td>Extension (A) (T)</td>
<td>2</td>
<td>High</td>
</tr>
<tr>
<td></td>
<td>Extension (G) (C)</td>
<td>2</td>
<td>High</td>
</tr>
<tr>
<td>Hybridization</td>
<td>Hyb (Low)</td>
<td>1</td>
<td>Low</td>
</tr>
<tr>
<td></td>
<td>Hyb (Medium)</td>
<td>1</td>
<td>Medium</td>
</tr>
<tr>
<td></td>
<td>Hyb (High)</td>
<td>1</td>
<td>High</td>
</tr>
<tr>
<td>Target removal</td>
<td>Target Removal 1, 2</td>
<td>2</td>
<td>Low</td>
</tr>
<tr>
<td>Bisulphite conversion I</td>
<td>BC conversion I (C1, C2, C3)</td>
<td>3</td>
<td>High</td>
</tr>
<tr>
<td></td>
<td>BC conversion I (U1, U2, U3)</td>
<td>3</td>
<td>Background</td>
</tr>
<tr>
<td></td>
<td>BC conversion I (C4, C5, C6)</td>
<td>3</td>
<td>High</td>
</tr>
<tr>
<td></td>
<td>BC conversion I (U4, U5, U6)</td>
<td>3</td>
<td>Background</td>
</tr>
<tr>
<td>Bisulphite conversion II</td>
<td>BC conversion II (1, 2, 3, 4)</td>
<td>4</td>
<td>High</td>
</tr>
<tr>
<td>Specificity I</td>
<td>GT mismatch 1, 2, 3 (PM)</td>
<td>3</td>
<td>High</td>
</tr>
<tr>
<td></td>
<td>GT mismatch 1, 2, 3 (MM)</td>
<td>3</td>
<td>Background</td>
</tr>
<tr>
<td>Specificity II</td>
<td>Specificity 1, 2, 3</td>
<td>3</td>
<td>High</td>
</tr>
<tr>
<td>Non-poly morphic</td>
<td>NP (A), (T)</td>
<td>2</td>
<td>High</td>
</tr>
<tr>
<td></td>
<td>NP (C), (G)</td>
<td>2</td>
<td>High</td>
</tr>
<tr>
<td>Negative</td>
<td>Average of all probes</td>
<td></td>
<td>Background</td>
</tr>
<tr>
<td></td>
<td>Standard deviation of all intensities</td>
<td>600</td>
<td>Background</td>
</tr>
</tbody>
</table>

Table 2.7: Expected intensities of control probes. Number and expected intensity shown for sample-independent and sample-dependent control probes. Adapted from (Illumina 2011).
The raw image data (idat) files produced during scanning of the BeadChips were input into R and using the program ‘minfi’ and quality control plots were subsequently generated (Hansen and Aryee 2013). These plots summarise the distribution of raw beta values for each sample, as well as the performance of the control probes, using beanplots and control strip plots, as a different way of visualising the performance of control probes and identifying outliers. Data was pre-processed without normalisation to create an object of class ‘Methylset’. A MethylSet is a class of object that contains the methylated and unmethylated values for each probe on the array for each sample, and from this M-values or beta-values can be extracted. Beta-values and M-values are both metrics that represent methylation levels (Du et al. 2010). Beta-values are a measurement of methylation between 0 and 1 which is calculated from the ratio between methylated signal intensity and overall intensity of that position (see Equation 1). M-values are the log2 ratio of the intensities of methylated and unmethylated signals, and can have infinite positive and negative values (see Equation 2). Quality control information was extracted from the MethylSet to generate a quality control plot defining good and poor quality samples by comparing the log transformed intensities from the methylated and unmethylated channels. If a sample did not surpass a cutoff of 10.5 it was considered a poor sample, as recommended by the package developers, and was excluded from further analysis (Hansen and Aryee 2014).
Equation 1: Calculation of Beta values of methylation from methylated and unmethylated probe intensities.
This equation defines the beta-value at a CpG site $i$ where $y_{i,methy}$ is the intensity measured by the methylated probe and $y_{i,unmethy}$ is the intensity measured by the unmethylated probe at position $i$. A constant offset, $\alpha$, is added to adjust data when both methylated and unmethylated probe intensities are low.

$$
Beta_i = \frac{\max(y_{i,methy}, 0)}{\max(y_{i,unmethy}, 0) + \max(y_{i,methy}, 0) + \alpha}
$$

Equation 2: Calculation of M-values of methylation from methylated and unmethylated probe intensities.
This equation defines the M-value at a CpG site $i$ where $y_{i,methy}$ is the intensity measured by the methylated probe and $y_{i,unmethy}$ is the intensity measured by the unmethylated probe at position $i$. The ratio of methylated and unmethylated intensities at site $i$ is log2 transformed. A constant offset, $\alpha$, is added to adjust data when both methylated and unmethylated probe intensities are low.

$$
M_i = \log_2 \left( \frac{\max(y_{i,methy}, 0) + \alpha}{\max(y_{i,unmethy}, 0) + \alpha} \right)
$$
2.3.1.3. Normalisation

SWAN was used to normalise technical differences between the two chemistries present on the array, i.e. Infinium I and Infinium II (section 2.2.5.1), without removing true biological differences that exist between samples (Maksimovic, Gordon, and Oshlack 2012). Over half of Infinium I probes are located within CpG islands whereas only approximately one fifth of Infinium II probes map to these regions. This is because, although Infinium II is the preferred chemistry, Infinium I probes produce more robust methylation estimates when >3 CpGs are present in the probe body— as can be the case in regions of high CpG density, i.e. in CpG islands. However, it has been previously noted that the distribution of beta-values are different for the two chemistries with Infinium II probes displaying a narrower range of values (Bibikova et al. 2011; Dedeurwaerder et al. 2011). SWAN allows Infinium I and II probes to be normalised together which has been demonstrated to improve correlation between technical replicates and between array data and pyrosequencing results (Maksimovic, Gordon, and Oshlack 2012). An example of distribution of beta values of one sample before and after SWAN is shown in Figure 2.12.

During SWAN, a subset of probes is used to determine an average quantile distribution, based on the number of CpG sites the probes contain, which acts as a surrogate for local CpG density (Maksimovic, Gordon, and Oshlack 2012). The subset from each probe type is then sorted by intensity then the mean for each ‘row’ of probe type intensities is calculated and assigned to that quantile. The remaining probes are then adjusted by extrapolating the distribution of the probe subset to Infinium I and Infinium II probes separately by linear interpolation.

.
Figure 2.12: Example of distribution of beta values of a sample before and after SWAN. Both plots are from the same sample, showing the distribution of Infinium I probe methylation (red dashed lines) and Infinium II probe methylation (blue dashed lines) and distribution of both probe types combined (black lines). The left plot shows the unnormalised (i.e. raw) data, while the right plot shows the data distribution following SWAN normalisation.
2.3.1.4. **Pruning unwanted probes**

Detection p-value was calculated to identify background signal levels of CpG probes. Probes that had a detection-p value of more than 0.01 were considered to have failed and were removed from analysis. Probes residing on the X (n=11,232) and Y (n=416) chromosome were removed, as were any probes that contained SNPs anywhere in the 50 base probe sequence, according to the Illumina annotation (Illumina 2013).

2.3.1.5. **Principal component analysis**

Principal component analysis was performed using the PCAmethods package (Stacklies et al. 2007). Each of the top 4 principal components was then correlated with possible confounding factors including technical measurements such as position on array, processing date and beadchip ID; and biological measurements such as gender, smoking status, concurrent treatment with DMARDs and age. The top 4 principal components were investigated as these were having the largest effect on the data, often representing more than 5% of the variability in the dataset.

2.3.1.6. **Adjusting for first principal component using ComBat**

Batch correction was applied using ‘ComBat’ in the ‘SVA’ package (Johnson, Li, and Rabinovic 2007; Leek et al. 2013). This method of adjusting batch effects using an empirical Bayes method and was developed for use in expression data, but it has been used successfully in DNA methylation microarray data (Sun et al. 2011).
2.3.1.7. Identifying differentially methylated positions

Differentially methylated positions between responders and non-responders to the anti-TNF drugs were identified using linear regression in the package ‘CpGassoc’ (Barfield, Conneely, and Kilaru 2012). A mixed effects model was used, with clinical covariates included as fixed covariates. In the individual drug analyses (ie etanercept or adalimumab), methylation array was included as a random effect while in the combined analysis, position on array was included as a random effect. The significance of associations was calculated in two ways, using a false discovery rate (FDR) using the Benjamini-Hochberg method (Benjamini and Yekutieli 2001), and using the Holm method to calculate Bonferroni significance (Holm 1979). Bonferroni was used as it was the most stringent measure of significance, and FDR was used to provide a lower threshold of significance.

2.3.1.8. Repeat of analysis to include SNP-containing probes

The analysis described in sections 2.3.1.5-2.3.1.7 was repeated with SNP-containing probes preserved in the dataset during QC.
2.3.1.9. **Estimation of cell composition**

In order to determine whether cell composition was affecting the results, as can happen when using a heterogeneous sample such as whole blood, the cell composition was estimated using the Houseman regression calibration algorithm (Houseman et al. 2012). This was performed using the estimateCellCounts function in minfi, using the RGChannelSet object as input. The RGChannelSet is an object which represents the raw intensities of the red and green channel set in two matrices. This can be preprocessed into a MethylSet using the array manifest from Illumina (Hansen and Aryee 2014). The function combines the DNA methylation data produced during the EWAS of treatment response with a reference flow sorted dataset in the Bioconductor package FlowSorted.Blood.450k (Jaffe 2014). This reference dataset consists of 6 individual cell types, sourced from a study that generated genome-wide DNA methylation data (using the 27k array) of seven isolated cell populations in 6 Swedish males (Reinius et al. 2012). The cell populations in this dataset were CD4+ T-cells, CD8+ T-cells, CD56+ NK-cells, CD19+ B-cells, CD14+ monocytes, and granulocytes. The reference data and etanercept response data were quantile normalised together to reduce batch effects between datasets. Probes within the flow sorted dataset were then selected to be used in composition estimation of the response dataset. The mean of each cell type in each response group was then calculated from which differences between groups could be calculated. A Welch two sample t-test was performed to measure differences between the response groups.

The analysis described in section 2.3.1.7 was then repeated using the estimated cell counts as covariates in the mixed effects association model to determine if adjustment for cell composition was necessary in samples and if it qualitatively changed results.
2.3.1.10. **Power calculation**

A power calculation based on the etanercept response EWAS was performed in STATA statistical software (StataCorp 2013) based on standard unpaired t-tests. These assume a normal distribution of the data and equal variances between groups. Power to detect differences in methylation was calculated at a significance level of 0.05 for 8%, 10% or 12% variance in methylation.

2.3.1.11. **Validation of epigenome-wide association studies with pyrosequencing**

Methylation values from pyrosequencing assays were calculated for each validation position of interest using the PyroMark Q24 Software version 2.0.6. This software was used to generate a “CpG Full Report” for each assay, from which percentage methylation was recorded. Samples which were considered to have ‘failed’ (due to insufficient signal) were excluded from analysis, as were any assays that appeared to have low signal quality across all samples (i.e. samples that were assigned the quality measure of ‘check’ by the PyroMark Q24 Software).

The methylation levels at each CpG site of interest were compared using STATA statistical software (StataCorp 2013). The β-values generated from the HumanMethylation450 BeadChip data were compared with percentage methylation as calculated by the PyroMark Q24 software version 2.0.6. The Spearman’s rank correlation coefficient (Spearman’s rho) was used to determine the correlation between methylation levels as measured on the HumanMethylation450 BeadChip and pyrosequencing results at the CpG site of interest for each individual. Spearman’s rho is a nonparametric statistical test for association between values; a Spearman’s rho of 1 or -1 represents perfect correlation, while 0 represents a lack of correlation.
2.3.2. Epigenome-wide association study of rheumatoid arthritis discordant monozygotic twins

2.3.2.1. Overview

In analysis of twin studies, it is important to consider the relatedness of twins during analysis as some statistical packages and algorithms are not compatible with twin analysis. In this study, quality control was performed in minfi, data was then SWAN normalised before probes that failed detection were removed along with probes that map to the X or Y chromosomes. Principal component analysis was performed to identify any confounding effect of patient characteristics or batch effects. These were then adjusted for in the mixed effects model to identify differentially methylated positions as fixed covariates. In the model, family was included as a random covariate to correct for family structure of MZ twins. A summary of the analysis pipeline is shown below in Figure 2.13.
Figure 2.13: Summary of analysis pipeline for identification of methylation differences in RA discordant monozygotic twins.
2.3.2.2.  *Quality control*

Quality control was performed using quality control plots generated in minfi to determine if any samples were outliers. These included kernel density plots of beta values for each sample and a plot comparing the log transformed median methylated and unmethylated intensities with a quality cutoff of 10.5 (as described in section 2.3.1.2).

2.3.2.3.  *Normalisation*

SWAN normalisation was performed to reduce the effects of differences in Infinium chemistry between type one and type two probes (described in more detail in section 2.3.1.3).

2.3.2.4.  *Pruning unwanted probes*

Probes that had a detection p-value of greater than 0.01 were removed, as were probes that resided in the X (n=11,232) and Y (n=416) chromosomes. Due to the identical genetic sequence between monozygotic twins, SNP-containing probes were not removed from analysis.

2.3.2.5.  *Principal components analysis*

PCA was used to identify the covariates in the samples that were having the biggest confounding effect on the data, as described in section 2.3.1.5.
2.3.2.6. Identifying differentially methylated positions

Differentially methylated positions were identified using β-values of methylation to identify differences between healthy and RA-affected twins using the package CpGassoc (Barfield, Conneely, and Kilaru 2012). In a mixed effects model, family was included as a random effect while the first and second principal components (as identified during PCA) were included in the model as fixed effects. The principal components were included as the first principal component, which represented a large proportion of the variability in the dataset, and could not be attributed to a single characteristic or batch effect.

2.3.2.7. Estimation of cell composition of whole blood in rheumatoid arthritis discordant twins

The cell composition of RA-affected and non-RA twins was estimated using the Houseman method (described in detail in section 2.3.1.9) (Houseman et al. 2012). This was performed using the estimateCellCounts function in minfi, using the RGChannelSet as input. Using a reference dataset (Jaffe 2014), the proportion of CD4+ T-cells, CD8+ T-cells, CD56+ NK-cells, CD19+ B-cells, CD14+ monocytes, neutrophils, and eosinophils was estimated using the regression calibration algorithm. The mean proportion of each cell type in each group (RA-affected twins and non-RA twins) was calculated and a Welch two sample T-test was performed to measure differences in cell composition between groups.

Differentially methylated positions were then identified with adjustment for cell composition. Analysis was performed as described in sections 2.3.2.2-2.3.2.6, but in the final mixed effects model, measurements of each cell type were included as fixed covariates along with the first and second principal component.
2.3.2.8. Validation of twin study using PBMC samples from rheumatoid arthritis discordant monozygotic twins

Twin data generated from PBMC samples were analysed separately from the whole blood samples described above. Analysis of these samples was performed as described in sections 2.3.2.3-2.3.2.7.
2.3.3. Investigation into effects of cryopreservation of cells on DNA methylation

2.3.3.1. Overview

Ten blood samples (taken in duplicate from five individuals) were processed as described in section 2.2.1 and the resultant PBMCs were used in the investigation of the effects of cryopreservation on DNA methylation levels. CD3+CD4+ T-cells were sorted from each PBMC sample as detailed in section 2.2.1.4. One sample from each individual (n=5) was cryopreserved prior to FACS separation (section 2.2.1), while for the other samples (n=5), CD3+CD4+ T-cells were extracted from the fresh sample. DNA samples were prepared for array hybridisation (as described in sections 2.2.5.2-2.2.5.11) and data quality control was performed as described in section 2.3.1.2. Differences in global levels of DNA methylation between cryopreserved and fresh samples was assessed by Pearson's product-moment correlation. Linear regression using the CpGassoc package (section 2.3.1.7) was then used to assess differential methylation between cryopreserved and fresh samples for each probe to identify specific changes in DNA methylation following cryopreservation of cells.

To investigate whether the process of FACS sorting alone altered DNA methylation, DNA methylation levels of two PBMC samples were compared as described above, with and without exposure to FACS.
2.3.3.2. Quality control for all samples

Quality control was performed together on all samples involved in this sub-study (n=12). Kernal density plots were generated in minfi (Hansen and Aryee 2013) and median log transformed signals from the methylated and unmethylated channels were plotted. In this quality control plot, a quality cut-off of 10.5 was used to identify poorly performing samples.

2.3.3.3. Epigenome-wide association study of CD3+CD4+ T-cells separated from cryopreserved versus fresh PBMCs

2.3.3.3.1. Normalisation

Samples were SWAN normalised (described in more detail in section 2.3.1.3).

2.3.3.3.2. Pruning unwanted probes

Probes that had a detection p-value of 0.01 or greater were removed along with probes mapping to the X and Y chromosomes.

2.3.3.3.3. Principal components analysis

Principal components analysis was performed using the PCAmethods package in R (Stacklies et al. 2007). Each principal component was then correlated with volunteer data (age, gender and individual ID number), possible batch effects (position on array) and the measurement of interest (whether a sample was cryopreserved or not).
2.3.3.4. **Identification of differentially methylated positions**

Differentially methylated positions were identified between samples that had been cryopreserved and those that had not using the CpGassoc package in R (Barfield, Conneely, and Kilaru 2012).

2.3.3.5. **Correlation of global DNA methylation levels**

The SWAN normalised pruned data sets for each pair of samples were correlated to measure the concordance of DNA methylation levels within each pair of samples. The Pearson's product-moment correlation was used to determine the similarity of distribution between cryopreserved and fresh samples from each individual.

2.3.3.4. **Epigenome-wide association study of PBMCs exposed to FACS versus not exposed to FACS**

Analysis was performed as described in section 2.3.3.3, except for identification of DMPs, which was not performed as there were only two samples included in analysis.
3. Results
3.1. Quality control of bisulphite conversion and validation of pyrosequencing technology with Sanger sequencing

3.1.1. Sanger sequencing of DAPK1

Pyrosequencing is a well-recognised method used to perform targeted DNA methylation studies; however, the technology was new to the lab at the time I started my PhD. In order to ensure that results generated using the Pyrosequencing platform were robust, the first step was to compare findings with the method of Sanger sequencing. Therefore, the DAPK1 region, recommended by Zymo for bisulphite conversion quality control, was assessed using both platforms to ensure that bisulphite conversion was complete and could be accurately measured using the pyrosequencing platform. The sequence was interpreted by comparing the Sanger trace with the template sequence of the region. The efficiency of bisulphite conversion was measured by assessing whether all the non-CpG cytosine bases had converted to a thymine residue, indicating complete bisulphite conversion: C residues outside of a CpG context are not expected to be methylated so a ‘T’ peak in the Sanger sequencing trace would be expected at that location in the sequence, while incomplete conversion would produce a double peak corresponding to the presence of both ‘C’ and ‘T’ alleles at the locus in question. If bisulphite conversion occurred to completion, then the proportion of methylation of each CpG site can be estimated approximately from the comparative size of the ‘C’ (blue) and ‘T’ (red) peaks (Figure 3.1). Due to variations in the start site of the Sanger sequencing and disruptions in the end of the sequence caused by a mononucleotide run (see section 3.1.1.1), a 50bp region in the middle of the sequencing results was selected to focus on for qualitative investigation of methylation.

The Sanger sequencing indicated that the DNA was fully bisulphite converted. This is evident as the non-CpG cytosine residues were completely converted to thymine
residues, producing strong peaks at these locations. The 50bp sequence contained eight non-CpG cytosine bases, and in all samples these positions produced a single ‘T’ peak indicating complete conversion. Concordance between samples was very good (Figure 3.1). This shows that the bisulphite conversion has been completed consistently across samples.

The 50bp sequence also contains six CpG sites. Across all three biological samples, these were all found to be unmethylated and fully converted, producing ‘T’ peaks. Both technical replicates of the methylated DNA standard (Figure 3.1 A and B) showed all CpG sites were methylated and remained cytosine residues, as expected. This shows that the bisulphite conversion protocol does not allow overconversion of bases, which can happen with over exposure to sodium bisulphite, causing conversion of methylated cytosine to uracil. Analysis of the technical replicates of the unmethylated DNA standard (Figure 3.1 C and D) revealed that, of the six CpG sites, one was methylated and was read in the Sanger trace as a ‘C’, while the remaining CpG sites had double peaks showing both C and T residues are present at these loci. This indicates partial methylation of the CpGs in the non-methylated DNA control.
Figure 3.1: Sanger sequencing traces for 24bp section of DAPK1. This region is approximately 50bp from the start of the sequenced product (this varies as sequencing trace begins at slightly different loci in different samples). Different coloured traces correspond to different bases (red: T, green: A, black: G and blue: C). The trace shows that C residues outside of CpG sites have all been converted to T residues, indicated with asterisks (*). This suggests bisulphite conversion has happened to completion. CpG sites are indicated with a red cross (×), which depending on methylation status should be either a C residue (methylated) or a T residue (unmethylated).
3.1.1.1. Reduced sequence complexity

Due to the conversion of many unmethylated cytosine bases to uracil (and post-PCR, to thymine) during bisulphite conversion, the DNA has reduced sequence complexity, as it is mostly made up of just three bases (A, G and T). This creates a problem during primer design, as primer sequences are much less likely to be specific to a certain region and are more likely to anneal in multiple locations throughout the genome. Thus, primers designed for bisulphite converted DNA must be longer than normal, must not contain CpG sites and any non-CpG cytosine residues must be replaced by T residues. This reduced complexity also creates problems during sequencing, as there is a greater potential for long runs of mono-nucleotide bases. During sequencing, long stretches of the same base can cause polymerase ‘slipping’, whereby the template can loop and rehybridize. This can cause the generation of different sized sequencing products, which can then cause a mixed signal in the trace following the mono-nucleotide run. In the sequenced amplicon, bisulphite conversion produced a run of 16 T’s followed by one G then another 6 T’s starting 140bp into the sequence (Figure 3.2). This produced sequence slippage in the sequencing trace for all samples, making all subsequent sequencing results unintelligible.

![Figure 3.2: Sanger sequencing trace showing long mononucleotide run causing sequence slippage and subsequent mixed signal.](image)
3.1.2. **Pyrosequencing of DAPK1_3 and DAPK1_5**

The region of *DAPK1* which was Sanger sequenced was subsequently analysed using pyrosequencing. Due to the length of the region, it was split into two assays (DAPK1_3 and DAPK1_5). The DAPK1_3 assay contained 17 CpG sites, while the DAPK1_5 assay contained 18 CpG sites.

Each DNA sample (n=5) was bisulphite converted in duplicate (i.e. two technical replicates for each sample), and the similarity in results between technical replicates for each sample demonstrated the reliability of both bisulphite conversion and the pyrosequencing methods. The similarity between technical replicates was assessed by comparing the percentage methylation at each CpG site, from within the targeted region. All methylation measurements were within 3% of each other in technical replicates of biological samples, with most loci being either the same (37 of 53 sites measured in DAPK1_5 across three samples and 33 of 51 sites measured in DAPK1_3) or just 1% different (16 of 53 sites in DAPK1_5 and 16 of 51 sites in DAPK1_3).

The biological DNA samples had very low levels of methylation ranging between 0-10% at each CpG site, and displayed very good concordance between technical replicates (example pyrogram shown in Figure 3.3). In all three biological samples across both assays, all technical replicate values were within 2% of each other.

The non-methylated control DNA had a higher than expected and very variable level of methylation across the assays ranging between 4-27% in the DAPK1_3 assay and 5-55% in the DAPK1_5 assay. Four sites in the DAPK1_5 assay showed greater than 50% methylation. This is consistent with results from Sanger sequencing of the same region which also showed partial methylation at the same CpG sites. The technical replicates of the non-methylated control DNA correlated well, with all CpG sites in the DAPK1_3 assay having methylation levels within 2%
of each other, and 15 of 18 sites in the DAPK1_5 assay. The methylated control DNA however produced less concordant results in both the DAPK1_3 and DAPK1_5 experiments.

In the DAPK1_3 assay, there was poor concordance of methylation percentages at CpG sites between technical replicates of the methylated control DNA. This could be due to the very large peak produced at the start of the run for this sample at the first dATP dispensation (Figure 3.4), which could have disrupted the measurements for the rest of the bases.

While these results indicate complete bisulphite conversion and strong concordance between technical replicates, the assays in all samples were found to have poor quality, indicated by the yellow and red boxes around the methylation measurements (Figure 3.3). This could be due to the large size of the assays, indicating that in future, assays should be designed to be much smaller (less than 100bp).

Bisulphite conversion quality control dispensations (indicated by the pale orange highlighted base in Figure 3.3) were used to measure efficiency of bisulphite conversion. These quality control dispensations measure the quantity of cytosine and thymine residues at non-CpG cytosines, which indicates whether bisulphite conversion has been efficient. All samples in both assays showed no cytosine residues were present, indicating complete bisulphite conversion.
Figure 3.3: Example pyrosequencing trace of DAPK1_5 assay of a biological DNA sample. Pyrograms showing technical replicates of DNA sample bisulphite converted and pyrosequenced separately. Highlighted regions indicate a CpG site, and the percentage at the top of these regions shows the C content and therefore amount of methylation in each CpG site. The sequence below the trace indicates the bases which were added in the pyrosequencing reaction, while the peak height indicates the amount of uptake of each base, i.e. how many of each base is in that place in the sequence.

Figure 3.4: Pyrogram indicating higher than expected peak height in assay. The subtle grey block histogram behind the peaks indicates the expected peak height at each base. The circled region indicated the peak area outside the histogram. This image is an example from a biological sample sequenced using the DAPK1_3 assay.
3.1.3. **Pyrosequencing of DAPK1_1**

Due to the poor quality of results from the DAPK1_3 and DAPK1_5 assays, bisulphite conversion and pyrosequencing reliability were measured in another region of the *DAPK1* gene (DAPK1_1). This assay was performed using pre-designed primers from Qiagen to measure methylation in each of the five DNA samples described in section 2.2.6.7 which were bisulphite converted in duplicate.

Interestingly, the methylated control DNA again showed high levels of methylation which ranged between 74-98% across the region. This is consistent with results seen in the Sanger sequencing and pyrosequencing of the other region of *DAPK1* investigated (see sections 3.1.1 and 3.1.2).

The methylation values between technical replicates of all samples were concordant to within 1-2% (see Figure 3.5 and Figure 3.6), indicating reliable bisulphite conversion and showing the integrity and accuracy of pyrosequencing results.
Figure 3.5: Pyrosequencing results from DAPK1_1 assay of methylated control DNA. Pyrograms showing replicates of methylated control DNA sample bisulphite converted and pyrosequenced separately using the DAPK1_1 pre-designed assay from Qiagen. Highlighted regions indicate a CpG site, and the percentage at the top of these regions shows the C content and therefore amount of methylation in each CpG site. The sequence below the trace indicates the bases which were added in the pyrosequencing reaction, while the peak height indicates the amount of uptake of each base, ie how many of each base is in that place in the sequence.
Figure 3.6: Example of pyrosequencing results from DAPK1_1 assay of a biological DNA sample. Highlighted regions indicate a CpG site, and the percentage at the top of these regions shows the C content and therefore amount of methylation in each CpG site. The sequence below the trace indicates the bases which were added in the pyrosequencing reaction, while the peak height indicates the amount of uptake of each base, i.e., how many of each base is in that place in the sequence. The concordance between peak height and percentage methylation at each CpG site between samples shows that bisulphite conversion and pyrosequencing are reliable and generate reproducible results.
3.1.4. **Summary of bisulphite conversion quality control experiment results**

In summary, this quality control experiment has shown that Sanger sequencing is not an appropriate technique for the accurate measurement of DNA methylation, due to the reduced sequence complexity of bisulphite converted DNA. The long mononucleotide runs seen in this experiment should be predicted from reference sequences and avoided during assay design. Poor quality of the DAPK1_3 and DAPK1_5 assays may suggest that future amplicons should be shorter, and consistently aberrant results indicate that the unmethylated DNA standard is unreliable and should not be used as a control DNA sample. Close concordance between technical replicates indicates the robustness of bisulphite conversion and pyrosequencing techniques for use in investigation of DNA methylation.
3.2. **Epigenome-wide association studies of anti-TNF treatment response**

3.2.1. **Epigenome-wide association study of etanercept response**

3.2.1.1. **Patient characteristics**

Patients were selected from the BRAGGSS cohort. At the start of the current study 890 patients had been recruited into BRAGGSS, 573 of which had reached 3-month follow up and were eligible for treatment response studies. 208 of these patients had been treated with etanercept and of these, 40 patients achieved clinical remission and were classified as good responders while 51 patients did not respond to therapy. From these, 36 good responders and 36 poor responders were selected, based on DNA quality (see section 2.2.3) and where possible, were matched for baseline characteristics. The baseline characteristics of these patients are summarised in Table 3.1 below. None of the baseline characteristics were significantly different between response groups, however the mean age was lower in responders compared with non-responders (p=0.07).

<table>
<thead>
<tr>
<th></th>
<th>Responders (n=36)</th>
<th>Non-Responders (n=36)</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age, mean (SD)</td>
<td>54.6 (11.4)</td>
<td>59.9 (12.2)</td>
<td>0.07</td>
</tr>
<tr>
<td>Female, n (%)</td>
<td>28 (78)</td>
<td>29 (81)</td>
<td>0.78</td>
</tr>
<tr>
<td>Smoking status</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Current, n (%)</td>
<td>5 (13.9)</td>
<td>4 (11.1)</td>
<td>0.84</td>
</tr>
<tr>
<td>Past, n (%)</td>
<td>14 (38.8)</td>
<td>11 (30.6)</td>
<td></td>
</tr>
<tr>
<td>Never, n (%)</td>
<td>11 (30.6)</td>
<td>12 (33.3)</td>
<td></td>
</tr>
<tr>
<td>Smoking data not available, n (%)</td>
<td>6 (16.7)</td>
<td>9 (25)</td>
<td></td>
</tr>
<tr>
<td>DAS28, mean (SD)</td>
<td>5.9 (1.2)</td>
<td>5.8 (0.76)</td>
<td>0.56</td>
</tr>
<tr>
<td>*ACPA positive, n (%)</td>
<td>25 (69.5)</td>
<td>20 (57.14)</td>
<td>0.53</td>
</tr>
<tr>
<td>Concurrent DMARD therapy, n (%)</td>
<td>31 (86)</td>
<td>32 (88)</td>
<td>0.72</td>
</tr>
<tr>
<td>**On methotrexate as DMARD, n (%)</td>
<td>21 (78)</td>
<td>25 (78)</td>
<td>0.97</td>
</tr>
</tbody>
</table>

Table 3.1: Patient characteristics for epigenome-wide association study of etanercept response.
*data for ACPA status available for 36 responders and 35 non-responders, **data for specific DMARD available for 27 responders and 32 non-responders
3.2.1.2. **Analysis of HumanMethylation450 BeadChip data**

3.2.1.2.1. **Quality control of array data**

The summary results for the sample-dependent and sample-independent control probes present on the BeadChip can be accessed through the control dashboard in the GenomeStudio software. These data were used to assess the performance of each stage in the preparation of the BeadChips. The sample-independent controls (n=13) assess the quality of staining, extension, hybridisation and target removal while sample-dependent controls (n=629) assess the quality of the bisulphite conversion of the DNA (this is measured on type 1 and type 2 probes separately), specificity of probes, non-polymorphic DNA and also includes negative controls (control probes described in more detail in section 2.3.1.2). The number of control probes on the array varies between 1 and 600 depending on the control probe in question. One sample from a poor responder, BIOP006001, was considered an outlier on the sample-dependent control probes, including the bisulphite conversion, specificity and non-polymorphic controls (example control plot is shown in Figure 3.7) indicating there was a problem with the DNA sample used for that individual, possibly due to incomplete bisulphite conversion.
Figure 3.7: Example of control dashboard produced in GenomeStudio for bisulphite conversion control probes targeting type 2 probes. One sample appears to be a distinct outlier from the total 72 samples shown, indicating a sample-dependent problem.

A number of bioconductor packages including minfi (Aryee et al. 2014), PCAmethods (Stacklies et al. 2007), SVA (Leek et al. 2013) and cpg.assoc (Barfield, Conneely, and Kilaru 2012) were used to analyse the methylation data in the R software environment version 3.1.0. Sample BIOP006001, identified as an outlier in the GenomeStudio control dashboard, was again identified as such in the kernel density plot generated in minfi, as can be seen in Figure 3.8. Sample BIOP006001 was again identified as a poorly performing sample as it did not reach the threshold for median intensity in either the methylated or unmethylated channels.
(Figure 3.9). Due to the aberrant signals for sample BIOP006001 in both the assessment of the GenomeStudio control dashboard and the minfi quality control reports, this sample was removed from further analysis.

**Figure 3.8**: Quality control kernel density plot produced in minfi showing distribution of beta values split into good and poor responders to etanercept. One sample (BIOP006001) had a density distribution that deviated from the other samples in the experiment.
Figure 3.9: Quality control graph produced using quality control information extracted from the unnormalised methylset object in minfi. This plot shows the log transformed median intensity for methylated and unmethylated positions for each sample. One sample (BIOP006001) was classified as a ‘bad’ sample because it did not reach the threshold for median intensity in either the methylated or unmethylated channels.
3.2.1.2.2. **Normalisation**

Within minfi, Subset-quantile Within Array Normalisation (SWAN) was used to reduce technical variation within and between arrays (Maksimovic, Gordon, and Oshlack 2012).

3.2.1.2.3. **Pruning unwanted probes**

Detection p-values were used to identify failed probes, i.e. probes with an intensity signal that was not significantly higher than the background signal (p>0.01). In total, 37,737 probes failed detection and were excluded from further analysis.

Probes containing SNPs within the probe sequence may display differential binding efficiency in different individuals due to genetic variants affecting the binding of probes to the DNA, as has been observed in gene expression arrays using probes of the same length (Benovoy, Kwan, and Majewski 2008). This effect could cause a genetic difference between groups to appear as a false positive methylation difference. For this reason, probes containing SNPs were excluded (n=148,966).

Following exclusion of failed probes, gender probes and SNP-containing probes, 318,488 probes remained available for further analysis.
3.2.1.2.4. **Principal components analysis**

The principal components analysis was performed in R using the ‘base’ and ‘PCAmethods’ packages using the singular value decomposition (SVD) method on the SWAN normalised pruned dataset. Eigenvalues were calculated and the variance in the data captured for each loading are shown in the plot of eigenvalue structure (Figure 3.10) and tabulated in Table 3.2. Each of the first four principal components was correlated with patient characteristics such as gender, age at baseline, smoking status, treatment with concurrent DMARDs and baseline DAS28 score, as well as with possible causes of batch effect such as position of sample on the BeadChip and which BeadChip the sample was hybridised to. The first four principal components were investigated as they each explained more than 5% of the variance in the dataset (Figure 3.10). The strongest correlation with the first principal component was position on the BeadChip (p=3.5E-9, Table 3.3). The BeadChip which the sample was hybridised to was strongly correlated with the fourth principal component (p=1.2E-4, Table 3.3). Gender was also correlated with the fourth principal component, but to a lesser degree than BeadChip (p=0.04, Table 3.3). No other measured patient characteristics were significantly correlated with the first four principal components.
Figure 3.10: Graph indicating the variability in the dataset which is explained by each principal component. The principal component index is indicated on the x-axis while the y-axis shows the variance in the dataset explained by each loading. Each of the first 4 loadings explain more than 5% of the variance and each of the first two loadings explain more than 10%.

Table 3.2: Amount of variance in the data explained by each principal component loading. Cumulatively the first ten principal components explain 52% of the variability in the data. $R^2$ shows the amount of variance explained by each individual component while cumulative $R^2$ shows the total variance explained by all components up until that component.
<table>
<thead>
<tr>
<th></th>
<th>PC1</th>
<th>PC2</th>
<th>PC3</th>
<th>PC4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Position on array</td>
<td>3.49E-9</td>
<td>0.09</td>
<td>7.08E-4</td>
<td>0.49</td>
</tr>
<tr>
<td>BeadChip</td>
<td>0.19</td>
<td>0.46</td>
<td>3.82E-4</td>
<td>1.18E-4</td>
</tr>
<tr>
<td>Gender</td>
<td>0.53</td>
<td>0.80</td>
<td>0.46</td>
<td>0.04</td>
</tr>
<tr>
<td>Age at baseline</td>
<td>0.68</td>
<td>0.58</td>
<td>0.50</td>
<td>0.80</td>
</tr>
<tr>
<td>Smoking status</td>
<td>0.64</td>
<td>0.47</td>
<td>0.59</td>
<td>0.72</td>
</tr>
<tr>
<td>Treatment with concurrent DMARDs</td>
<td>0.47</td>
<td>0.72</td>
<td>0.83</td>
<td>0.83</td>
</tr>
<tr>
<td>Baseline DAS score</td>
<td>0.17</td>
<td>0.99</td>
<td>0.61</td>
<td>0.30</td>
</tr>
<tr>
<td>Etanercept response</td>
<td>0.30</td>
<td>0.43</td>
<td>0.08</td>
<td>0.95</td>
</tr>
</tbody>
</table>

Table 3.3: Correlation of first four principal component loadings with batch and patient specific measurements. Values indicate p-value for association between that principal component and a particular patient characteristic.
3.2.1.2.5. **Adjusting for batch effect using ComBat**

The location in which the sample was positioned on the BeadChip was identified as a batch effect following principal components (PC) analysis. This unwanted variation was removed from the methylation data using the ComBat method in the bioconductor package SVA (Johnson, Li, and Rabinovic 2007). PC analysis performed following ComBat found that the effect of sample position had been removed. The correlation between sample position and the first PC pre-ComBat and post-ComBat can be seen in Figure 3.11 and Figure 3.12, respectively.
Figure 3.11: Relationship between position on array of sample and the first principal component in the data, prior to ComBat adjustment. There is a clear correlation between position on array and PC1 ($p=3.5 \times 10^{-9}$), indicating the importance of adjusting this in the data to prevent confounding.

Figure 3.12: Relationship between position of sample on array and the first principal component after ComBat adjustment. The correlation between PC1 and position on array has been removed ($p=0.54$) showing that ComBat analysis has successfully adjusted this confounding effect.
3.2.1.2.6. Identifying differentially methylated positions

Differentially methylated positions were identified using the CpG assoc bioconductor package (Barfield, Conneely, and Kilaru 2012). In this package mixed effects models can be specified. Many other packages (e.g. minfi, methylumi, Limma and lumi) which are commonly used to identify differentially methylated positions do not allow the inclusion of BeadChip or batch as a random intercept. In the mixed model, patient characteristics that are known to influence response (Hyrich et al. 2006; Kristensen et al. 2008) e.g. gender, baseline DAS28 score and treatment with concurrent DMARDs were included as fixed effects. The BeadChip ID was included as a random covariate. This was included because of the significant association between the fourth principal component and BeadChip ID.

Four CpG sites were identified as being differentially methylated between responders and non-responders to etanercept in 71 patients at the false discovery rate threshold of 5%. Of these, two CpG sites passed the more stringent Bonferroni correction. The results of the association test are summarised in Figure 3.13 and the 4 loci most strongly correlated with etanercept response are presented in Table 3.4. The most differentially methylated position cg04857395 (p=1.46E-8) maps to a CpG island in exon 7 of the gene body of the low density lipoprotein receptor-related protein associated protein 1 (LRPAP1) gene on chromosome 4 (Table 3.4 and Figure 3.14). This gene encodes a chaperone of low density lipoprotein receptor-related protein 1 (LRP1), which is important in TGFβ regulation. The second most differentially methylated position cg16426293 (p=1.31E-7) maps to an intergenic region on chromosome 17, 2068 base pairs away from the nearest gene, zinc finger protein 385C (ZNF385C) (Table 3.4). The third most differentially methylated position (p=2.22E-7), cg03277049, overlaps with the non-coding RNA, LINC00886 on chromosome 3. The fourth CpG that surpassed the 5% false discovery rate was cg14862806 (p=4.43E-7) on chromosome 17. This CpG does not map proximally to
any known genes (physical position 21356311). All four differentially methylated
CpG sites presented in Table 3.4 had higher levels of methylation in non-
responders to etanercept, compared with responders. The top DMP had a
difference of 9% methylation between response groups, while the second, third and
fourth most differentially methylated positions had differences of 6%, 6% and 3%
respectively (Table 3.4).

572 base pairs from the most differentially methylated position at the *LRPAP1*
locus, and in the same CpG island, another differentially methylated position was observed
albeit with less statistical significance (cg00356499, p=1.75E-4) as shown in Figure
3.14. Further, three additional probes mapping to the same CpG island were
present on the methylation array but were excluded during analysis due to the
presence of SNPs in the probe sequences, i.e. cg26401028, cg27321949 and
cg08297640. The analysis was therefore repeated to include these probes (see
section 3.2.1.2.7).
Figure 3.13: Manhattan plot of the relationship between response to etanercept and DNA methylation. Each point on the Manhattan Plot represents one CpG site that was included in the analysis. The y-axis show the $-\log_{10}(p$-values) for the association of each CpG site with methylation while the x-axis shows the chromosomal distribution of CpG sites. The dashed line represents the false discovery rate of 5% which four CpG sites surpassed, while the solid line represents the more stringent Bonferroni significance which two CpG sites surpassed.
<table>
<thead>
<tr>
<th>DMP</th>
<th>P-value of difference</th>
<th>Mean (SD) β-values in responders</th>
<th>Mean (SD) β-values in non-responders</th>
<th>Chromosome: physical position (annotation)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cg04857395</td>
<td>1.46E-8</td>
<td>0.72 (0.06)</td>
<td>0.81 (0.06)</td>
<td>Chr.4: 3516637</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>(In the gene body of LRPAP1)</td>
</tr>
<tr>
<td>Cg16426293</td>
<td>1.31E-7</td>
<td>0.48 (0.05)</td>
<td>0.54 (0.04)</td>
<td>Chr.17: 40192112</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>(2068bp from ZNF385C)</td>
</tr>
<tr>
<td>Cg03277049</td>
<td>2.22E-7</td>
<td>0.31 (0.05)</td>
<td>0.37 (0.04)</td>
<td>Chr.3: 156534076</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>(In LINC00886 non-coding RNA)</td>
</tr>
<tr>
<td>Cg14862806</td>
<td>4.43E-7</td>
<td>0.35 (0.02)</td>
<td>0.38 (0.03)</td>
<td>Chr.17: 21356311</td>
</tr>
</tbody>
</table>

Table 3.4: Differentially methylated positions between responders and non-responders to etanercept.
The four differentially methylated positions (DMPs) shown were associated with etanercept response at the 5% false discovery rate threshold. The p-value of the association between response to etanercept and DNA methylation at the position targeted by that probe is shown. The mean β-value of methylation at that CpG site in responders and non-responders to etanercept is presented along with the chromosomal position of each CpG site and proximity to genes.

Figure 3.14: Position of CpG island in relation to the LRPAP1 gene and position of the differentially methylated CpG sites. Probes that were excluded from analysis due to the presence of SNPs in the probe sequence are indicated with hollow circles.
3.2.1.2.7.  Identifying differentially methylated positions with inclusion of SNP-containing probes

In the previously described analysis SNP-containing probes were removed to avoid potential confounding, including three CpG sites within the *LRPAP1* gene. Given the association with the gene, the analysis was repeated including the SNP-containing probes. Following this analysis, 9 CpG sites surpassed the false discovery rate of 5% while 2 surpassed the more stringent Bonferroni significance threshold (see Figure 3.15 and Table 3.5). The probe cg04857395 remained the most significant result when SNP-containing probes were included (Table 3.5 and Figure 3.16). The other CpG site that passed Bonferroni significance (cg26401028) was also in the gene body of *LRPAP1* as was another site that passed the 5% FDR threshold (cg27321949). The remaining sites within the CpG island were also found to be differentially methylated albeit with less significant p-values (Figure 3.16).
Figure 3.15: Manhattan Plot of analysis of etanercept EWAS response results, with SNP-containing probes included. In this analysis, 9 CpG sites passed a false discovery rate of 0.05 (dashed line), two of which also surpassed the Bonferroni significance threshold (solid line).
Table 3.5: Differentially methylated positions which passed false discovery rate of 0.05 in etanercept response analysis in which SNP-containing probes were included. The table shows the top nine differentially methylated positions, the p-value of each sites association with etanercept response, and the chromosomal position of each site, with associated gene annotation. (*indicates SNP-containing probes)

<table>
<thead>
<tr>
<th>Ranking of DMP</th>
<th>DMP</th>
<th>P-value of difference</th>
<th>Chromosome: physical position (annotation)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>cg04857395</td>
<td>1.29E-8</td>
<td>Chr.4: 3516637 (In the gene body of LRPAP1)</td>
</tr>
<tr>
<td>2</td>
<td>cg26401028*</td>
<td>1.64E-8</td>
<td>Chr.4: 3516534 (In the gene body of LRPAP1)</td>
</tr>
<tr>
<td>3</td>
<td>cg16426293</td>
<td>1.32E-7</td>
<td>Chr.17: 40192112</td>
</tr>
<tr>
<td>4</td>
<td>cg03277049</td>
<td>2.29E-7</td>
<td>Chr.3: 156534076</td>
</tr>
<tr>
<td>5</td>
<td>cg14862806</td>
<td>4.16E-7</td>
<td>Chr.17: 21356311</td>
</tr>
<tr>
<td>6</td>
<td>cg27321949*</td>
<td>4.57E-7</td>
<td>Chr.4: 3516692 (In gene body of LRPAP1)</td>
</tr>
<tr>
<td>7</td>
<td>cg17162319*</td>
<td>8.24E-7</td>
<td>Chr.7: 57927783</td>
</tr>
<tr>
<td>8</td>
<td>cg01449144</td>
<td>9.35E-7</td>
<td>Chr.8: 49782592</td>
</tr>
<tr>
<td>9</td>
<td>cg12226028</td>
<td>9.49E-7</td>
<td>Chr.15: 76484473 (In gene body of C15orf27)</td>
</tr>
</tbody>
</table>
Figure 3.16: Position of CpG island in relation to *LRPAP1* gene and position of differentially methylated CpG sites across island, including previously excluded SNP-containing probes. When previously excluded SNP-containing probes were analysed, there was significant differential methylation at all CpG sites present on the array for this CpG island and its’ N-shore. This indicates that the differential methylation is more likely to be robust.
3.2.1.2.8. *Estimation of cell composition*

Whole blood is a heterogeneous sample with multiple cell types, each defined by differences in mitotically heritable epigenetic changes that underpin transcriptional programming of gene expression within cell types (Houseman *et al.* 2012; Reinius *et al.* 2012). Differences in DNA methylation levels between groups of samples could arise, therefore, as a result of different cellular compositions of the comparison groups. Hence, in the current study, cell composition was estimated in the whole blood samples using the Houseman method (Houseman *et al.* 2012) as described in section 2.3.1.9. There were no significant differences (p<0.05) between any of the six cell types estimated (i.e. CD8+ T-cells, CD4+ T-cells, CD56+ NK-cells, CD19+ B-cells, CD14+ monocytes and granulocytes) between responders and non-responders. As expected, granulocytes were the most abundant cell type in both responders and non responders with a proportion of 58-60% (Table 3.6). NK cells were the least abundant, making up 4% in both groups. The cells with the greatest difference between groups were the granulocytes, and the smallest difference between groups was in NK cells.

<table>
<thead>
<tr>
<th>Cell type</th>
<th>Proportion in responders</th>
<th>Proportion in non-responders</th>
<th>Difference</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>CD8T</td>
<td>5.09</td>
<td>5.58</td>
<td>0.49</td>
<td>0.61</td>
</tr>
<tr>
<td>CD4T</td>
<td>18.80</td>
<td>20.24</td>
<td>1.44</td>
<td>0.48</td>
</tr>
<tr>
<td>NK cells</td>
<td>4.49</td>
<td>4.29</td>
<td>0.20</td>
<td>0.81</td>
</tr>
<tr>
<td>B cells</td>
<td>4.99</td>
<td>5.84</td>
<td>0.85</td>
<td>0.28</td>
</tr>
<tr>
<td>Monocytes</td>
<td>8.80</td>
<td>8.58</td>
<td>0.22</td>
<td>0.78</td>
</tr>
<tr>
<td>Granulocytes</td>
<td>60.21</td>
<td>57.58</td>
<td>2.63</td>
<td>0.35</td>
</tr>
</tbody>
</table>

Table 3.6: Cell composition estimations for responders and non-responders to etanercept. Quantity of each cell type in response groups was estimated from the HumanMethylation450 BeadChip data using the Houseman reference based method. There was no significant difference in cell composition between response groups. P-value calculated using the Welch two sample t-test.
3.2.1.2.9. **Identifying differentially methylated positions with adjustment for cell composition**

Despite the lack of significant differences in cell composition between responders and non-responders to etanercept, it was important to ensure that adjusting for the cell counts would not significantly alter the results and conclusions. The analysis (as described in section 2.3.1.7) was, therefore, repeated with the addition of the individual cell count for each cell type (CD8T, CD4T, NK cells, B cells, monocytes and granulocytes) included as fixed covariates in the linear regression model. There were small changes in the results, i.e. 14 of the top 20 differentially methylated positions remained in the top 20 following adjustment for cell type (top 20 differentially methylated positions in data adjusted and unadjusted for cell composition shown in Table 3.7). Interestingly, the top three most differentially methylated CpG sites remain the same whether the data is adjusted for cell composition or not. When the data was adjusted for cell composition, the 5 most differentially methylated probes (Table 3.7) passed the false discovery rate of 5%, and cg04857395 (LRPAP1) remained significant after accounting for multiple testing using the Bonferroni method (see Figure 3.17).
<table>
<thead>
<tr>
<th>Rank of</th>
<th>Data unadjusted for cell composition</th>
<th>Data adjusted for cell composition</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>CpG site</td>
<td>p-value</td>
</tr>
<tr>
<td>DMP</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>cg04857395</td>
<td>1.46E-08</td>
</tr>
<tr>
<td>2</td>
<td>cg16426293</td>
<td>1.31E-07</td>
</tr>
<tr>
<td>3</td>
<td>cg03277049</td>
<td>2.22E-07</td>
</tr>
<tr>
<td>4</td>
<td>cg14862806</td>
<td>4.43E-07</td>
</tr>
<tr>
<td>5</td>
<td>cg12226028</td>
<td>9.36E-07</td>
</tr>
<tr>
<td>6</td>
<td>cg01449144</td>
<td>1.14E-06</td>
</tr>
<tr>
<td>7</td>
<td>cg14001858</td>
<td>1.35E-06</td>
</tr>
<tr>
<td>8</td>
<td>cg24411302</td>
<td>1.57E-06</td>
</tr>
<tr>
<td>9</td>
<td>cg03760457</td>
<td>2.36E-06</td>
</tr>
<tr>
<td>10</td>
<td>cg16824477</td>
<td>2.44E-06</td>
</tr>
<tr>
<td>11</td>
<td>cg05792368</td>
<td>2.50E-06</td>
</tr>
</tbody>
</table>
Table 3.7: Comparison of differentially methylated positions associated with response to etanercept in data adjusted for cell composition versus data not adjusted for cell composition.
Table shows the top twenty differentially methylated positions, the p-value of each sites association with etanercept response, and the associated gene annotation. Four probes passed false discovery rate of 5% when cell composition was not adjusted for, while five probes passed this threshold following adjustment for cell composition (indicated in bold). The lists of top twenty differentially methylated positions were very similar wether cell composition is adjusted for or not, and notably the three most significant probes in both analyses were the same.

<table>
<thead>
<tr>
<th></th>
<th>cg16677058</th>
<th>2.63E-06</th>
<th>PCP2</th>
<th>cg02746514</th>
<th>3.72E-06</th>
<th>FOXK1</th>
<th>0.76 (0.02)</th>
<th>0.78 (0.02)</th>
</tr>
</thead>
<tbody>
<tr>
<td>12</td>
<td>cg04988476</td>
<td>2.91E-06</td>
<td>FGF12</td>
<td>cg14862806</td>
<td>3.91E-06</td>
<td>Intergenic</td>
<td>0.35 (0.02)</td>
<td>0.38 (0.03)</td>
</tr>
<tr>
<td>13</td>
<td>cg20066782</td>
<td>3.13E-06</td>
<td>FBXL13;LRRC17</td>
<td>cg05792368</td>
<td>4.05E-06</td>
<td>Intergenic</td>
<td>0.13 (0.02)</td>
<td>0.15 (0.03)</td>
</tr>
<tr>
<td>14</td>
<td>cg01350680</td>
<td>3.16E-06</td>
<td>Intergenic</td>
<td>cg12226028</td>
<td>4.57E-06</td>
<td>C15orf27</td>
<td>0.30 (0.07)</td>
<td>0.41 (0.11)</td>
</tr>
<tr>
<td>15</td>
<td>cg17195635</td>
<td>3.16E-06</td>
<td>Intergenic</td>
<td>cg01350680</td>
<td>4.78E-06</td>
<td>Intergenic</td>
<td>0.28 (0.03)</td>
<td>0.32 (0.05)</td>
</tr>
<tr>
<td>16</td>
<td>cg15647481</td>
<td>3.28E-06</td>
<td>MTMR9L</td>
<td>cg05081953</td>
<td>5.28E-06</td>
<td>ERMM</td>
<td>0.49 (0.08)</td>
<td>0.41 (0.06)</td>
</tr>
<tr>
<td>17</td>
<td>cg09198277</td>
<td>5.00E-06</td>
<td>NOTCH4</td>
<td>cg16677058</td>
<td>5.66E-06</td>
<td>PCP2</td>
<td>0.93 (0.01)</td>
<td>0.94 (0.01)</td>
</tr>
<tr>
<td>18</td>
<td>cg26499055</td>
<td>5.06E-06</td>
<td>SIM1</td>
<td>cg17832858</td>
<td>6.32E-06</td>
<td>CYTH3</td>
<td>0.89 (0.05)</td>
<td>0.92 (0.02)</td>
</tr>
<tr>
<td>19</td>
<td>cg27400447</td>
<td>5.17E-06</td>
<td>CLYBL</td>
<td>cg17195635</td>
<td>6.47E-06</td>
<td>Intergenic</td>
<td>0.90 (0.03)</td>
<td>0.92 (0.02)</td>
</tr>
</tbody>
</table>
After adjustment for cell composition, five probes reached a FDR 5% (dashed line). One of these probes also passed the more stringent Bonferroni significance threshold (solid line).
3.2.1.2.10.  *Power calculation*

Power to detect differences in methylation was calculated for the 71 samples in this study using the STATA statistical software. Power was calculated at a significance level of 5% for 8%, 10% and 12% variance in methylation. There was more than 90% power to detect a 10% absolute change in methylation between the two groups, under the assumption that there is 12% or less variance in methylation.

<table>
<thead>
<tr>
<th>Absolute change in methylation</th>
<th>Power for p=0.05</th>
</tr>
</thead>
<tbody>
<tr>
<td>Two group comparison (n=35 vs. N=36)</td>
<td></td>
</tr>
<tr>
<td>SD of methylation = 8%</td>
<td></td>
</tr>
<tr>
<td>3%</td>
<td>0.3521</td>
</tr>
<tr>
<td>4%</td>
<td>0.5582</td>
</tr>
<tr>
<td>5%</td>
<td>0.7495</td>
</tr>
<tr>
<td>10%</td>
<td>0.9995</td>
</tr>
<tr>
<td>SD of methylation = 10%</td>
<td></td>
</tr>
<tr>
<td>3%</td>
<td>0.2438</td>
</tr>
<tr>
<td>4%</td>
<td>0.3918</td>
</tr>
<tr>
<td>5%</td>
<td>0.5582</td>
</tr>
<tr>
<td>10%</td>
<td>0.9879</td>
</tr>
<tr>
<td>SD of methylation = 12%</td>
<td></td>
</tr>
<tr>
<td>3%</td>
<td>0.1835</td>
</tr>
<tr>
<td>4%</td>
<td>0.2896</td>
</tr>
<tr>
<td>5%</td>
<td>0.4190</td>
</tr>
<tr>
<td>10%</td>
<td>0.9395</td>
</tr>
</tbody>
</table>

Table 3.8: Power calculation for EWAS of etanercept response at a significance level of p=0.05.
3.2.1.3. Technical validation of etanercept response EWAS using pyrosequencing

To provide a technical validation of the EWAS results, the top five most differentially methylated probes from the main analysis (excluding SNP-containing probes and not adjusting for cell composition) were assayed using pyrosequencing in 39 samples. The assays used in this validation are described in Table 3.9 below. The Spearman’s rank correlation coefficients for these probes are presented in the Table 3.10 below. The methylation values as measured by the two platforms correlated for cg04857395, cg16426293, cg14862806 and cg12226028, with Spearman’s rho values of 0.87, 0.86, 0.90 and 0.79 respectively (an example of the correlation of results for cg04857395 is shown in Figure 3.18 below). The CpG site identified by the probe ID cg03277049, correlated less well with a Spearman’s rho of 0.57.
Table 3.9: Pyrosequencing assays for validation of top 5 most differentially methylated positions from etanercept response epigenome-wide association study. All sequences are shown in the 5’ to 3’ direction.
<table>
<thead>
<tr>
<th>Rank of CpG site</th>
<th>CpG site</th>
<th>Spearman's rank</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>cg04857395</td>
<td>0.87</td>
</tr>
<tr>
<td>2</td>
<td>cg16426293</td>
<td>0.86</td>
</tr>
<tr>
<td>3</td>
<td>cg03277049</td>
<td>0.57</td>
</tr>
<tr>
<td>4</td>
<td>cg14862806</td>
<td>0.90</td>
</tr>
<tr>
<td>5</td>
<td>cg12226028</td>
<td>0.79</td>
</tr>
</tbody>
</table>

Table 3.10: Correlation of HumanMethylation450 BeadChip results with pyrosequencing results from the top most differentially methylated positions. The correlation of the BeadChip data with pyrosequencing data is indicated by the Spearman's rank correlation coefficient for each CpG site.

Figure 3.18: Correlation of the percentage methylation between HumanMethylation450 BeadChip (x-axis) and pyrosequencing data for the most differentially methylated position, cg04857395.
3.2.1.4. Summary of etanercept EWAS results

This epigenome-wide association study of response to etanercept identified four differentially methylated positions which reached study-wide significance when the data was analysed without SNP-containing probes and was not adjusted for cell composition. Cg04857395 was the most differentially methylated position throughout all analyses undertaken, with a 9% difference in methylation between responders and non-responders. This probe maps to the LRPAP1 gene in a CpG island which contains four additional differentially methylated positions, making it a differentially methylated region. No significant differences in cell composition between responders and non-responders to etanercept were identified. A technical validation of the five most differentially methylated positions using pyrosequencing showed good correlation for four sites, but a poor correlation for cg03277049.
3.2.2. **Epigenome-wide association study of adalimumab response**

3.2.2.1. **Patient characteristics**

Patients were again selected from the BRAGGSS cohort. At the time of selection, the study had recruited 1,416 patients, 784 of whom had reached 3-month follow up and were eligible for treatment response analysis. 265 patients were treated with adalimumab, 72 of which were considered good responders and 45 considered non-responders based on EULAR classification criteria. Once patients who did not have a good quality DNA sample were removed, 48 good responders and 27 non-responders remained. The mean age at baseline was higher in non-responders than responders, as was the proportion of females, so the three youngest male responders were removed leaving the final 72 patients for the study. The characteristics of these patients are detailed in Table 3.11 below. There were no significant differences between responders and non-responders for age, gender, smoking status, baseline DAS28 or concurrent methotrexate use, however there was a significant difference in concurrent DMARD therapy (p=0.0081).

<table>
<thead>
<tr>
<th></th>
<th>Responders (n=45)</th>
<th>Non-Responders (n=27)</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Age, mean (SD)</strong></td>
<td>53.28 (11.97)</td>
<td>53.20 (14.35)</td>
<td>0.98</td>
</tr>
<tr>
<td><strong>Female, n (%)</strong></td>
<td>36 (80)</td>
<td>23 (85)</td>
<td>0.58</td>
</tr>
<tr>
<td><strong>Smoking status</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Current, n (%)</td>
<td>5 (11)</td>
<td>5 (19)</td>
<td></td>
</tr>
<tr>
<td>Past, n (%)</td>
<td>19 (42)</td>
<td>6 (22)</td>
<td>0.65</td>
</tr>
<tr>
<td>Never, n (%)</td>
<td>15 (33)</td>
<td>12 (44)</td>
<td></td>
</tr>
<tr>
<td><strong>Smoking data not available, n (%)</strong></td>
<td>6 (13)</td>
<td>4 (15)</td>
<td></td>
</tr>
<tr>
<td><strong>DAS28, mean (SD)</strong></td>
<td>5.65 (0.78)</td>
<td>5.79 (1.00)</td>
<td>0.49</td>
</tr>
<tr>
<td><strong>Concurrent DMARD therapy, n (%)</strong></td>
<td>43 (96)</td>
<td>20 (74)</td>
<td>0.0081</td>
</tr>
<tr>
<td><em>On methotrexate as DMARD, n (%)</em></td>
<td>31 (72.09)</td>
<td>17 (62.96)</td>
<td>0.52</td>
</tr>
</tbody>
</table>

*data for specific DMARD available for 43 responders and all non-responders

Table 3.11: Patient characteristics for epigenome-wide association study of adalimumab response.
3.2.2.2. **Analysis of HumanMethylation450 BeadChip data**

3.2.2.2.1. **Quality control of array data**

The GenomeStudio control dashboard showed that all samples had the expected signal intensities of control probes (see Figure 3.19 for an example of the green extension control plot, expected signal intensities are listed in section 2.3.1.2).

![Graph showing sample-dependent or sample-independent problems](image)

**Figure 3.19**: Example of GenomeStudio Control plots used to assess performance of arrays and identify sample-dependent or sample-independent problems. This plot is an example of the green extension control probes showing the intensity of the control probes in each of the 72 samples in the study. The intensity of the probes is shown on the y-axis while the sample identifier is on the x-axis. The intensity of the samples tends to cluster into groups depending on which BeadChip the sample was on. The samples show a change in intensity according to their location on the BeadChip, showing the importance of adjusting for position on array during analysis.
The QC report produced in minfi did not identify any samples as outliers when the distribution of beta values was assessed (Figure 3.20). The QC plot produced in minfi using QC information extracted from the un-normalised methylset compared the median intensity of the methylated and unmethylated channels, however, identified one sample that did not pass the threshold of 10.5 (Figure 3.21). Since this sample was right on the threshold, and because the study was already underpowered, it was included in further analysis.

Figure 3.20: Density plot of beta values from each sample coloured by good and poor response to adalimumab. Green and orange lines indicate the beta distribution of good and poor responders to adalimumab, respectively. No outliers are visible from this plot and the expected bimodal distribution of methylation is clear indicating samples are of good quality.
Figure 3.21: Quality control graph produced using QC information extracted from the unnormalised methylset object in minfi. One sample in this plot is considered a ‘bad sample’ because the median methylated and unmethylated intensity signals for that sample did not surpass the cutoff of 10.5.
3.2.2.2. **Normalisation**

Samples were normalised using SWAN as described in section 3.2.1.2.2.

3.2.2.3. **Pruning unwanted probes**

Detection p-values were used to identify probes which did not have a strong enough signal. In total, 46,771 probes with a detection p-value greater than 0.01 were excluded (a full description of detection p-values can be found in section 2.3.1.4).

Following removal of failed probes, probes on the sex chromosomes and SNP-containing probes, 313,558 probes were retained for further analysis.

3.2.2.4. **Principal components analysis**

Principal components analysis was performed to identify batch effects which may be affecting the data. The amount of variability explained by each PC is shown in Figure 3.22. Each of the first four PCs were correlated with patient characteristics and technical batch effects including position on BeadChip and BeadChip number, and with patient characteristics such as gender, age, smoking status and treatment with concurrent DMARDs (Table 3.12). The first PC was strongly associated with position on BeadChip (p=2.88E-11) and showed a less significant association with gender (p=0.019) and baseline DAS28 (p=0.044). BeadChip number was strongly associated with the second PC (p=1.21E-7). Age at baseline (p=0.009) and position on BeadChip (p=0.042) showed less significant associations with the third PC whereas smoking status and concurrent DMARD treatment were not associated with any of the first four PCs.
Figure 3.22: Variability of methylation data explained by each PC in the adalimumab treatment response study. Each point represents a principal component loading, with points above the solid line indicating loadings that explain more than 10% of variability in the dataset.
<table>
<thead>
<tr>
<th></th>
<th>PC1</th>
<th>PC2</th>
<th>PC3</th>
<th>PC4</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Position on BeadChip</strong></td>
<td>2.88E-11</td>
<td>0.8641</td>
<td>0.04153</td>
<td>0.01583</td>
</tr>
<tr>
<td><strong>BeadChip number</strong></td>
<td>0.6701</td>
<td>1.21E-7</td>
<td>0.1459</td>
<td>0.514</td>
</tr>
<tr>
<td><strong>Gender</strong></td>
<td>0.01858</td>
<td>0.7286</td>
<td>0.8745</td>
<td>0.334</td>
</tr>
<tr>
<td><strong>Age at baseline</strong></td>
<td>0.05313</td>
<td>0.7157</td>
<td>0.009412</td>
<td>0.211</td>
</tr>
<tr>
<td><strong>Smoking status</strong></td>
<td>0.1693</td>
<td>0.1644</td>
<td>0.7982</td>
<td>0.1222</td>
</tr>
<tr>
<td><strong>Treatment with concurrent dmards</strong></td>
<td>0.8404</td>
<td>0.2668</td>
<td>0.8531</td>
<td>0.4993</td>
</tr>
<tr>
<td><strong>Baseline DAS28 score</strong></td>
<td>0.04457</td>
<td>0.3534</td>
<td>0.799</td>
<td>0.643</td>
</tr>
<tr>
<td><strong>Response to adalimumab</strong></td>
<td>0.7933</td>
<td>0.6674</td>
<td>0.1957</td>
<td>0.7195</td>
</tr>
</tbody>
</table>

Table 3.12: Correlation of first four PCs with batch and patient characteristics in analysis of methylation in adalimumab treatment response. Values displayed are p-values of associations between the PCs and batch or patient characteristics. The first PC was strongly associated with position on array while the second principal component was associated with array number.
3.2.2.5. Adjusting for biggest batch effect using ComBat

Position on BeadChip was adjusted using ComBat (Johnson, Li, and Rabinovic 2007) as described in section 3.2.1.2.5.

3.2.2.6. Identification of differentially methylated positions

Differentially methylated positions were identified using a mixed effects model in the CpGassoc package. Gender, baseline DAS28 and concurrent treatment with DMARDs were included as fixed covariates as they have been previously identified as affecting treatment response to anti-TNF biologics (Emery and Dorner 2011; Hyrich et al. 2006; Kristensen et al. 2008). BeadChip number was included as a random covariate. The top 20 most associated positions are shown in Table 3.13. While no CpG sites passed FDR 0.05 (Figure 3.23), several of the top most differentially methylated positions map to interesting genes or regions.
Figure 3.23: Manhattan plot of association between CpG sites and response to adalimumab. Each CpG site on the array is represented as a point on the manhattan plot, distributed by chromosome (x-axis). The y-axis represents the $-\log^{10}$ adjusted p-values of the association between each CpG site and adalimumab response. No CpG sites passed the FDR of 0.05.
<table>
<thead>
<tr>
<th>Rank</th>
<th>DMP</th>
<th>P-value of association</th>
<th>Mean (SD) β-values in responders</th>
<th>Mean (SD) β-values in non-responders</th>
<th>Chromosome: physical position (associated gene)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>cg10165222</td>
<td>3.59E-07</td>
<td>0.08 (0.01)</td>
<td>0.06 (0.01)</td>
<td>Chr10:119134612 (PDZD8)</td>
</tr>
<tr>
<td>2</td>
<td>cg09121920</td>
<td>3.59E-06</td>
<td>0.95 (0.01)</td>
<td>0.94 (0.01)</td>
<td>Chr16:397672 (AXIN1)</td>
</tr>
<tr>
<td>3</td>
<td>cg09044186</td>
<td>4.99E-06</td>
<td>0.94 (0.02)</td>
<td>0.92 (0.04)</td>
<td>Chr11:116661388 (APOA5)</td>
</tr>
<tr>
<td>4</td>
<td>cg05857826</td>
<td>5.70E-06</td>
<td>0.02 (0.003)</td>
<td>0.02 (0.003)</td>
<td>Chr20:30795350</td>
</tr>
<tr>
<td>5</td>
<td>cg26668837</td>
<td>6.09E-06</td>
<td>0.05 (0.01)</td>
<td>0.06 (0.01)</td>
<td>Chr11:8615356 (STK33)</td>
</tr>
<tr>
<td>6</td>
<td>cg22859422</td>
<td>1.23E-05</td>
<td>0.94 (0.01)</td>
<td>0.93 (0.01)</td>
<td>Chr2:241463367 (ANKMY1)</td>
</tr>
<tr>
<td>7</td>
<td>cg06641357</td>
<td>1.36E-05</td>
<td>0.92 (0.02)</td>
<td>0.91 (0.01)</td>
<td>Chr16:29197753 (intergenic)</td>
</tr>
<tr>
<td>8</td>
<td>cg15188808</td>
<td>1.41E-05</td>
<td>0.97 (0.01)</td>
<td>0.98 (0.01)</td>
<td>Chr7:150094161 (ZNF775)</td>
</tr>
<tr>
<td>9</td>
<td>cg08294017</td>
<td>1.85E-05</td>
<td>0.91 (0.01)</td>
<td>0.90 (0.01)</td>
<td>Chr2:220432822 (OBSL1)</td>
</tr>
<tr>
<td>10</td>
<td>cg16748413</td>
<td>2.38E-05</td>
<td>0.11 (0.02)</td>
<td>0.12 (0.01)</td>
<td>Chr8:82543431 (intergenic)</td>
</tr>
<tr>
<td>11</td>
<td>cg16636355</td>
<td>2.41E-05</td>
<td>0.15 (0.02)</td>
<td>0.20 (0.06)</td>
<td>Chr10:22764665 (intergenic)</td>
</tr>
<tr>
<td>12</td>
<td>cg09430647</td>
<td>3.30E-05</td>
<td>0.82 (0.03)</td>
<td>0.78 (0.04)</td>
<td>Chr10:106247914 (intergenic)</td>
</tr>
<tr>
<td>13</td>
<td>cg16782848</td>
<td>3.80E-05</td>
<td>0.05 (0.01)</td>
<td>0.06 (0.01)</td>
<td>Chr15:83479213 (WHAMM)</td>
</tr>
<tr>
<td>14</td>
<td>cg10018615</td>
<td>4.03E-05</td>
<td>0.14 (0.03)</td>
<td>0.11 (0.03)</td>
<td>Chr2:956401381 (intergenic)</td>
</tr>
<tr>
<td>15</td>
<td>cg15591276</td>
<td>4.49E-05</td>
<td>0.31 (0.06)</td>
<td>0.38 (0.07)</td>
<td>Chr7:1265851 (intergenic)</td>
</tr>
<tr>
<td>16</td>
<td>cg04989440</td>
<td>4.80E-05</td>
<td>0.07 (0.02)</td>
<td>0.09 (0.02)</td>
<td>Chr6:125684580 (intergenic)</td>
</tr>
<tr>
<td>17</td>
<td>cg25217317</td>
<td>4.86E-05</td>
<td>0.31 (0.03)</td>
<td>0.35 (0.05)</td>
<td>Chr1:235811994 (GNG4)</td>
</tr>
<tr>
<td>18</td>
<td>cg23601376</td>
<td>5.50E-05</td>
<td>0.07 (0.01)</td>
<td>0.08 (0.02)</td>
<td>Chr2:40678101 (SLC8Al)</td>
</tr>
<tr>
<td>19</td>
<td>cg26078407</td>
<td>5.76E-05</td>
<td>0.50 (0.02)</td>
<td>0.52 (0.02)</td>
<td>Chr2:240225062 (HDAC4)</td>
</tr>
<tr>
<td>20</td>
<td>cg04931493</td>
<td>6.17E-05</td>
<td>0.62 (0.05)</td>
<td>0.57 (0.06)</td>
<td>Chr16:3246099 (intergenic)</td>
</tr>
</tbody>
</table>

Table 3.13: Top 20 CpG sites at which methylation is associated with response to adalimumab (not adjusted for cell composition). The top twenty differentially methylated positions are shown, the p-value for the association with adalimumab response, mean beta-value of methylation in responder and non-responder groups and the chromosomal position of each site, with associated gene annotation also displayed. None of the sites passed false discovery rate of 5%.
3.2.2.2.7. Estimation of cell composition

Proportion of monocytes, granulocytes, CD8T, CD4T, NK and B cells were estimated using the Houseman method in minfi (Hansen and Aryee 2014; Houseman et al. 2012). There were no significant differences in cell proportions between responders and non-responders to adalimumab (Table 3.14). Cell compositions were very similar to those measured in etanercept patients (see section 3.2.1.2.8).

<table>
<thead>
<tr>
<th>Cell type</th>
<th>Proportion in responders</th>
<th>Proportion in non-responders</th>
<th>Difference</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>CD8T</td>
<td>5.38</td>
<td>6.43</td>
<td>1.05</td>
<td>0.41</td>
</tr>
<tr>
<td>CD4T</td>
<td>17.92</td>
<td>20.75</td>
<td>2.83</td>
<td>0.23</td>
</tr>
<tr>
<td>NK cells</td>
<td>3.76</td>
<td>3.65</td>
<td>0.11</td>
<td>0.91</td>
</tr>
<tr>
<td>B cells</td>
<td>4.16</td>
<td>4.73</td>
<td>0.57</td>
<td>0.57</td>
</tr>
<tr>
<td>Monocytes</td>
<td>9.84</td>
<td>9.38</td>
<td>0.46</td>
<td>0.71</td>
</tr>
<tr>
<td>Granulocytes</td>
<td>59.21</td>
<td>55.70</td>
<td>3.51</td>
<td>0.26</td>
</tr>
</tbody>
</table>

Table 3.14: Proportions of cells in responders and non-responders to adalimumab
3.2.2.2.8. Identifying differentially methylated positions with adjustment for cell composition

After adjustment for cell composition, the top most differentially methylated position from the unadjusted analysis, cg10165222, remained the most associated (see Figure 3.24 and Table 3.15), however, this site now passed both FDR 0.05 and bonferroni correction. Five of the top 20 most differentially methylated positions were the same with or without correction for cell composition (Table 3.15) showing that results were not altered qualitatively; however they were altered more than the etanercept DMPs.

Figure 3.24: Manhattan plot of association between CpG sites and response to adalimumab with adjustment for cell composition. 
-log p-values (y-axis) indicates significance of association of methylation with response to adalimumab, arranged by chromosome (x-axis). One CpG site passed both a FDR of 0.05 (solid line) and the more stringent Bonferroni significance (dashed line).
<table>
<thead>
<tr>
<th>Rank of DMP</th>
<th>Data unadjusted for cell composition</th>
<th>Data adjusted for cell composition</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>CpG site</td>
<td>p-value</td>
</tr>
<tr>
<td>1</td>
<td>cg10165222</td>
<td>3.59E-07</td>
</tr>
<tr>
<td>2</td>
<td>cg09121920</td>
<td>3.59E-06</td>
</tr>
<tr>
<td>3</td>
<td>cg09044186</td>
<td>4.99E-06</td>
</tr>
<tr>
<td>4</td>
<td>cg05857826</td>
<td>5.70E-06</td>
</tr>
<tr>
<td>5</td>
<td>cg26668837</td>
<td>6.09E-06</td>
</tr>
<tr>
<td>6</td>
<td>cg22859422</td>
<td>1.23E-05</td>
</tr>
<tr>
<td>7</td>
<td>cg06641357</td>
<td>1.36E-05</td>
</tr>
<tr>
<td>8</td>
<td>cg15188808</td>
<td>1.41E-05</td>
</tr>
<tr>
<td>9</td>
<td>cg08294017</td>
<td>1.85E-05</td>
</tr>
<tr>
<td>10</td>
<td>cg16748413</td>
<td>2.38E-05</td>
</tr>
</tbody>
</table>
Table 3.15: Top 20 most differentially methylated positions between responders and non-responders to adalimumab with and without adjustment for cell composition.

The top ten differentially methylated positions are shown, along with the p-value for the association with adalimumab treatment response, and the associated gene annotation. The only site that passed false discovery rate of 5% was cg10165222 (highlighted in bold) following adjustment for cell composition.

<table>
<thead>
<tr>
<th></th>
<th>cg</th>
<th>p-Value</th>
<th></th>
<th>cg</th>
<th>p-Value</th>
<th>Gene</th>
<th>p-Value</th>
<th>Gene</th>
</tr>
</thead>
<tbody>
<tr>
<td>11</td>
<td>cg16636355</td>
<td>2.41E-05</td>
<td>Intergenic</td>
<td>cg18124616</td>
<td>5.68E-05</td>
<td>USP1</td>
<td>0.07 (0.02)</td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>cg09430647</td>
<td>3.30E-05</td>
<td>Intergenic</td>
<td>cg25640176</td>
<td>6.24E-05</td>
<td>GFOD2</td>
<td>0.05 (0.01)</td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>cg16782848</td>
<td>3.80E-05</td>
<td>WHAMM</td>
<td>cg22355889</td>
<td>6.99E-05</td>
<td>ELMOD1; LOC643923</td>
<td>0.08 (0.07)</td>
<td></td>
</tr>
<tr>
<td>14</td>
<td>cg10018615</td>
<td>4.03E-05</td>
<td>Intergenic</td>
<td>cg21148836</td>
<td>7.84E-05</td>
<td>MEX3B</td>
<td>0.05 (0.01)</td>
<td></td>
</tr>
<tr>
<td>15</td>
<td>cg15591276</td>
<td>4.49E-05</td>
<td>Intergenic</td>
<td>cg09262504</td>
<td>8.40E-05</td>
<td>THSD4</td>
<td>0.91 (0.02)</td>
<td></td>
</tr>
<tr>
<td>16</td>
<td>cg04989440</td>
<td>4.80E-05</td>
<td>Intergenic</td>
<td>cg09378238</td>
<td>8.75E-05</td>
<td>CHN2</td>
<td>0.09 (0.01)</td>
<td></td>
</tr>
<tr>
<td>17</td>
<td>cg25217317</td>
<td>4.86E-05</td>
<td>GNG4</td>
<td>cg24917065</td>
<td>0.000101</td>
<td>SLC25A37</td>
<td>0.78 (0.04)</td>
<td></td>
</tr>
<tr>
<td>18</td>
<td>cg23601376</td>
<td>5.50E-05</td>
<td>SLC8A1</td>
<td>cg09177277</td>
<td>0.000101</td>
<td>DAPK2</td>
<td>0.10 (0.02)</td>
<td></td>
</tr>
<tr>
<td>19</td>
<td>cg26078407</td>
<td>5.76E-05</td>
<td>HDAC4</td>
<td>cg06405563</td>
<td>0.000105</td>
<td>Intergenic</td>
<td>0.49 (0.05)</td>
<td></td>
</tr>
<tr>
<td>20</td>
<td>cg04931493</td>
<td>6.17E-05</td>
<td>Intergenic</td>
<td>cg00910333</td>
<td>0.000116</td>
<td>TARBP2</td>
<td>0.02 (0.005)</td>
<td></td>
</tr>
</tbody>
</table>
3.2.2.3. Technical validation of adalimumab EWAS using pyrosequencing

The five most differentially methylated positions identified during EWAS of adalimumab response analysis (cg10165222, cg09121920, cg09044186, cg05857826 and cg26668837) were assessed using pyrosequencing to validate the BeadChip results in a subset of 24 samples using the assays shown in Table 3.16 (see Figure 3.25 for an example pyrogram). The results of pyrosequencing of cg10165222 were unreliable due to low signal of the assays and so these results were excluded from further analysis, as were measurements from any individual assays that failed (2 individuals in the assay targetting cg09044186 and 1 in cg26668837). When the four remaining pyrosequencing assays were compared with the EWAS data, the Spearman's rho for correlation of pyrosequencing and BeadChip results was 0.91 (Figure 3.26). While the assays correlated very well when interrogated together, individually they correlated less well (see Table 3.17). Cg09044186 and cg26668837 correlated relatively well with Spearman rho values of 0.67 and 0.56 respectively (an example plot of the correlation of cg09044186 is shown in Figure 3.27). Cg09121920 and cg05857826 had poor correlation, with Spearman rho values of 0.23 and 0.13 respectively. This poor correlation is likely to be due to the lack of variability of methylation (<1% in BeadChip measurements) in these samples at these sites, as measured by both the BeadChip and pyrosequencing (Figure 3.28).
Table 3.16: Pyrosequencing assays for validation of top 5 most differentially methylated positions from adalimumab response epigenome-wide association study
All sequences are shown in the 5’ to 3’ direction

<table>
<thead>
<tr>
<th>CpG of interest</th>
<th>Primer sequences</th>
<th>Number of CpG sites interrogated by assay</th>
<th>Primer set score</th>
<th>Size of PCR product (bp)</th>
</tr>
</thead>
<tbody>
<tr>
<td>cg10165222</td>
<td>GAGTAGGTTAGGTATTTGATGTAG CTCTCCCCCAATTCTTCTCTATAC TGATGTAGAGGAAGTTTT</td>
<td>11</td>
<td>82</td>
<td>111</td>
</tr>
<tr>
<td>cg09121920</td>
<td>AGGTGGTGGAGTTTTTTGAAT AAAAACACCTACATATATTCTCAT GGTATTGTTTTTTATTTTTT</td>
<td>4</td>
<td>89</td>
<td>122</td>
</tr>
<tr>
<td>cg09044186</td>
<td>ATGGGTGGAAAGTTTTTTGA AAAACACAAACCCATACTAA GGTGAAGAGTTTTTTGAA</td>
<td>6</td>
<td>83</td>
<td>114</td>
</tr>
<tr>
<td>cg05857826</td>
<td>ATATAGGTATGGGGTTAGGGA ACAATACACCAACCACAAAACCAGG GGGTGGAGATGAG</td>
<td>30</td>
<td>56</td>
<td>318</td>
</tr>
<tr>
<td>cg26668837</td>
<td>GTTTTAGATGGGGTGAGGTGA AAAACCCACCTTAATACCCCAATTCCCA TGGGGTGGAGGTGAG</td>
<td>4</td>
<td>94</td>
<td>105</td>
</tr>
</tbody>
</table>
Figure 3.25: Example of a program showing percentage methylation at each CpG site (highlighted within the blue bars) measured by the proportion of peak signal. The bisulphite conversion control (highlighted in yellow) had no peak indicating complete bisulphite conversion. The ratio of each peak signal indicates the number of bases of that kind that have been incorporated into the sequence, i.e. a peak of approximately 200 represents one base being incorporated while a peak of approximately 400 represents two bases.

Figure 3.26: Correlation of methylation values of four CpG sites measured by pyrosequencing (y-axis) versus HumanMethylation450 BeadChip signal (x-axis) in validation study.
<table>
<thead>
<tr>
<th>Rank of CpG site</th>
<th>CpG site</th>
<th>Spearman's rho</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>Cg09121920</td>
<td>0.23</td>
</tr>
<tr>
<td>3</td>
<td>Cg09044186</td>
<td>0.67</td>
</tr>
<tr>
<td>4</td>
<td>Cg05857826</td>
<td>0.13</td>
</tr>
<tr>
<td>5</td>
<td>Cg26668837</td>
<td>0.56</td>
</tr>
</tbody>
</table>

Table 3.17: Correlation of pyrosequencing with HumanMethylation450 BeadChip results for differentially methylated positions in Adalimumab response.

![Graph of cg09044186 DNA methylation levels as measured by pyrosequencing (y-axis) versus HumanMethylation450 BeadChip signal (x-axis).]

Figure 3.27: Correlation of cg09044186 DNA methylation levels as measured by pyrosequencing (y-axis) versus HumanMethylation450 BeadChip signal (x-axis).
Figure 3.28: Correlation of cg05857826 DNA methylation levels as measured by pyrosequencing (y-axis) versus HumanMethylation450 BeadChip signal (x-axis).
3.2.2.4. Summary of adalimumab EWAS results

In this epigenome-wide association study of response to adalimumab, I identified one CpG site which passed Bonferroni significance after adjustment of the data for cell composition. This CpG site, cg10165222, lies in the PDZD8 gene body near the transcription start site. When the data was unadjusted for cell composition, this site remained the most differentially methylated position but did not surpass the FDR of 5%. While small differences in methylation (1-3%) were detected between responders and non-responders to adalimumab, these are not large enough to be biologically significant. A technical validation of the top five most differentially methylated positions (one of which was excluded from analysis due to poor signal) from the analysis performed without adjustment for cell composition showed acceptable correlation for two sites, and poor correlation for two sites, which is likely to be due to lack of variability at these sites in the subset of samples investigated.
3.2.3. **Combined analysis of epigenome-wide association study of etanercept and adalimumab response**

It was important to analyse the individual drug classes separately, as described in sections 3.2.1 and 3.2.2, because etanercept and adalimumab inhibit TNF via different mechanisms (see section 1.2.3 for details). However, as both drugs are anti-TNFs, there may be overlap in loci that underpin response to the drug class as a whole. To investigate this, the data from both sub-studies were analysed together.

3.2.3.1. *Quality control*

The samples from both studies had already been checked for quality. Just one sample from the etanercept study, BIOP006001, was excluded from analysis (as described previously in 3.2.1.2.1) while all adalimumab samples were included. This resulted in a total of 143 individuals being included in the combined analysis, as summarised in Table 3.18.

<table>
<thead>
<tr>
<th></th>
<th>Responders</th>
<th>Non-responders</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Etanercept study</strong></td>
<td>36</td>
<td>35</td>
</tr>
<tr>
<td><strong>Adalimumab study</strong></td>
<td>45</td>
<td>27</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td>81</td>
<td>62</td>
</tr>
</tbody>
</table>

Table 3.18: Summary of number of patients included in combined analysis of etanercept and adalimumab data.
3.2.3.2. Normalisation

SWAN normalisation was performed on the data as described in section 3.2.1.2.2.

3.2.3.3. Pruning unwanted probes

In total, 84,508 probes had a detection p-value of more than 0.01 and were therefore removed. The probes mapping to the sex chromosomes were removed as were probes that contained SNPs in the probe sequence. This left 312,059 probes for further analysis.

3.2.3.4. Principal components analysis

Principal components analysis was performed to identify any batch or cohort characteristic which may be affecting the data (Table 3.19). The first PC was strongly associated with the date the BeadChips were processed (p<2.2E-16). Since the adalimumab and etanercept studies were prepared on separate dates (over a year apart), the date batch completely correlated with which anti-TNF therapy the patient was treated with. Therefore a caveat to adjusting for date batch using ComBat, is that differences between groups caused by differences in anti-TNFs will also be adjusted out. Position on array was associated with the second PC (p<2.2E-16) while array number was associated with the third PC, but to a lesser extent (p=1.85E-7). Age at baseline was slightly associated with the fourth PC (p=0.053).
<table>
<thead>
<tr>
<th></th>
<th>PC1</th>
<th>PC2</th>
<th>PC3</th>
<th>PC4</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Date batch</strong></td>
<td>&lt; 2.2E-16</td>
<td>0.005447</td>
<td>0.0009298</td>
<td>0.6335</td>
</tr>
<tr>
<td><strong>Position on array</strong></td>
<td>0.2504</td>
<td>&lt; 2.2E-16</td>
<td>0.1017</td>
<td>0.08745</td>
</tr>
<tr>
<td><strong>Array number</strong></td>
<td>2.622E-15</td>
<td>0.04257</td>
<td>1.85E-07</td>
<td>0.6513</td>
</tr>
<tr>
<td><strong>Gender</strong></td>
<td>0.2618</td>
<td>0.06395</td>
<td>0.5011</td>
<td>0.1969</td>
</tr>
<tr>
<td><strong>Age at baseline</strong></td>
<td>0.0056</td>
<td>0.2479</td>
<td>0.2089</td>
<td>0.05277</td>
</tr>
<tr>
<td><strong>Smoking status</strong></td>
<td>0.08894</td>
<td>0.0821</td>
<td>0.2916</td>
<td>0.3319</td>
</tr>
<tr>
<td><strong>Treatment with concurrent DMARDs</strong></td>
<td>0.8692</td>
<td>0.8407</td>
<td>0.6854</td>
<td>0.6732</td>
</tr>
<tr>
<td><strong>Which anti-TNF therapy</strong></td>
<td>&lt; 2.2E-16</td>
<td>0.000313</td>
<td>2.179E-05</td>
<td>0.239</td>
</tr>
<tr>
<td><strong>Baseline DAS</strong></td>
<td>0.9762</td>
<td>0.5112</td>
<td>0.09304</td>
<td>0.2565</td>
</tr>
</tbody>
</table>

Table 3.19: Associations between the first four principal components and possible batch effects or patient characteristics. Values shown are p-values for association between each principal component with each possible batch (date sample was processed, position on array and array number) or patient characteristic (gender, age at baseline, smoking status, concurrent treatment with DMARDs, which anti-TNF therapy the patient is treated with and baseline DAS). PC: principal component.
3.2.3.5. *Adjusting for batch effect using ComBat*

ComBat was used to adjust for effect of the date BeadChips were processed (as described in (Johnson, Li, and Rabinovic 2007)).

3.2.3.6. *Identification of differentially methylated positions*

Differentially methylated positions were identified using a mixed effects model in the CpG.assoc package (Barfield, Conneely, and Kilaru 2012). Age at baseline and smoking status were included as fixed effects while position on array was included as a random effect. Although no significant individual DMPs (FDR=0.05) were identified (Figure 3.29), a very interesting region was identified as being differentially methylated in the overlapping *CRYZ/TYW3* genes. Within a 700bp region 9 CpG sites are targeted by the BeadChip, 8 of which were differentially methylated between good and poor responders (p<1e-4) as shown in Figure 3.30, with the remaining probe, cg07399417, removed during QC. All 8 of the DMPs were among the top 40 most differentially methylated sites, and 3 were also in the top 5 (cg26690034, cg02709834 and cg26855724). Other DMPs were identified nearby to this region, including cg04250926 (p=2.9E-4) which lies 374bp upstream from cg00121533 and cg10128416 (p=1E-3) which lies 179bp upstream from cg21906852. This indicates that the region is quite extensively differentially methylated between responders and non-responders to anti-TNF therapies.
Figure 3.29: Manhattan plot of association between methylation levels and response to anti-TNF therapies. Y-axis indicates $-\log^{10}$ p-values of association between each CpG site methylation levels and response to the anti-TNF therapies, etanercept and adalimumab.
Figure 3.30: Differentially methylated positions in the overlapping CRYZ/TYW3 gene regions associated with response to anti-TNF therapies.

This diagram indicates the 9 CpG sites that lie within a 700bp region. All these CpG sites were found to be differentially methylated between responders and non-responders to anti-TNFs, with 9 being in the top 40 most differentially methylated positions, 3 of which were in the top 5.

3.2.3.7. Summary of combined analysis

In the combined analysis of etanercept and adalimumab, a differentially methylated region that spanned 700bp was identified that overlapped the transcription start sites of the CRYZ and TYW3 genes. This is a particularly interesting region, as it has previously been associated in genetic studies with circulating resistin levels, which are associated with inflammation, type-2 diabetes and cardiovascular disease (Qi et al. 2012). Most recently, the CRYZ/TYW3 gene was found to be differentially methylated in whole blood of RA discordant MZ twins (n=28 pairs) (Svendsen et al. 2014). While this is an interesting finding, these results should be treated with caution due to the large batch effect between the two studies. In future, meta-analysis of data may be more appropriate when combining datasets to prevent confounding by batch effects, and to allow identification of methylation differences which may be in opposing directions in different drug classes.
3.3. Epigenome-wide association study of rheumatoid arthritis discordant monozygotic twins

3.3.1. Patient characteristics

Patients were selected from the Rheumatoid Arthritis Twin Study in Manchester and from the TwinsUK cohort (Department of Twin Research and Genetic Epidemiology, King’s College London, UK). In the Rheumatoid Arthritis Twin Study, 77 monozygotic twin pairs were assessed by a rheumatologist and one twin was classified as having RA while the other twin was assessed not to have RA. Forty five pairs of twins had a good quality whole blood (WB) DNA sample available for analysis. Of the remaining 32 pairs of twins where a good quality whole blood DNA sample was not available, a PBMC sample was identified for 17 pairs from which DNA could be extracted. Seventeen twin pairs were eligible from TwinsUK and had a good quality WB DNA sample available for both twins. The studies from which samples were sourced and the type of samples used in the epigenome-wide association study are detailed in Table 3.20 below.

The PBMC samples were treated differently to the WB samples; the PBMC DNA was extracted from samples which had been cryopreserved for a long period of time and been exposed to freeze-thaw cycles, while DNA from WB was extracted from fresh samples. Because sample processing differences could result in batch effects and potentially cause confounding in the study, WB samples were analysed separately to PBMC samples. The characteristics of patients separated by RA status and sample type are shown below in Table 3.21. In total a WB sample was available for 62 RA discordant twin pairs, and a PBL sample was available for 17 RA discordant twin pairs. There were no significant differences in age or smoking status between groups in either WB (p=0.99 and 0.52, respectively) or PBMC samples (p=0.98 and 0.36, respectively).
<table>
<thead>
<tr>
<th>Study</th>
<th>Sample type</th>
<th>Number of RA discordant twin pairs</th>
</tr>
</thead>
<tbody>
<tr>
<td>RATS</td>
<td>Whole blood</td>
<td>45</td>
</tr>
<tr>
<td></td>
<td>PBMC</td>
<td>17</td>
</tr>
<tr>
<td>TwinsUK</td>
<td>Whole blood</td>
<td>17</td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td>79</td>
</tr>
</tbody>
</table>

Table 3.20: Summary of RA discordant twin pairs.

The table shows the study from which twin pairs originated and the sample type for the pairs. RATS= Rheumatoid arthritis twin study, PBMC= peripheral blood lymphocytes, RA= rheumatoid arthritis

<table>
<thead>
<tr>
<th></th>
<th>Whole blood (n= 124)</th>
<th>PBMC (n=34)</th>
</tr>
</thead>
<tbody>
<tr>
<td>RA status of twin</td>
<td>RA (n=62)</td>
<td>Healthy (n=62)</td>
</tr>
<tr>
<td>Age at sample collection in years, mean (SD)</td>
<td>54.57 (12.38)</td>
<td>54.60 (12.34)</td>
</tr>
<tr>
<td>Female, n (%)</td>
<td>51 (82)</td>
<td>51 (82)</td>
</tr>
<tr>
<td>Smoking status</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Current, n (%)</td>
<td>12 (19)</td>
<td>10 (16)</td>
</tr>
<tr>
<td>Past, n (%)</td>
<td>19 (31)</td>
<td>17 (27)</td>
</tr>
<tr>
<td>Never, n (%)</td>
<td>31 (50)</td>
<td>35 (56)</td>
</tr>
</tbody>
</table>

Table 3.21: Cohort characteristics for RA discordant twins with whole blood samples.

Rheumatoid arthritis (RA) status for each group of twins is shown stratified by the source of the sample, from whole blood or peripheral blood mononuclear cell (PBMC) samples. Mean and standard deviation (SD) of age is shown as well as number (n) and percentage (%) of patients who are female and who smoke currently, smoked in the past or never smoked. The p-value of differences between RA and healthy patients in each sample type is shown and was calculated using the kwallis test for equality of populations.
3.3.2. **Analysis of whole blood twin pairs**

3.3.2.1. *Quality Control*

Quality of samples was determined by quality control plots produced in minfi. Kernel density plots of samples allowed identification of outliers using the RGset object (Figure 3.31). A plot of log-transformed median intensity for methylated and unmethylated positions for each sample was produced from the Methylset object (see section 3.2.1.2.1 for more detail). In this plot, all samples surpassed the threshold for median intensity so were included in the analysis (Figure 3.32).

![Quality control kernel density plot of beta value distribution in RA discordant MZ twin samples. Each line indicates a sample involved in the twin study (RA and healthy co-twin).](image)

Figure 3.31: Quality control kernel density plot of beta value distribution in RA discordant MZ twin samples. Each line indicates a sample involved in the twin study (RA and healthy co-twin)
3.3.2.2. Normalisation

SWAN normalisation was performed on the data as described in section 3.2.1.2.2.

3.3.2.3. Pruning unwanted probes

In total 39,311 probes with a detection p-values of >0.01 were removed, and 11,648 probes residing on the X or Y chromosomes. This resulted in 470,523 probes for further analysis.
3.3.2.4. **Principal components analysis**

Principal components analysis was performed in R using the ‘base’ and ‘PCAmethods’ packages. The SVD method was used, and eigenvalues were calculated to measure how much variability in the dataset is accounted for by each PC (Figure 3.33, tabulated in Table 3.22). The first PC accounted for over 20% of the variability in the dataset, while the second accounted for 6%. These two PCs were compared to patient characteristics and possible causes of batch effects to identify the source of variability associated with that PC (Table 3.23). The first PC did not correlate directly with any of the batch or patient characteristics tested. The centre which the DNA sample was sourced from (i.e. TwinsUK or Arthritis Research UK centre for Genetics and Genomics) was strongly associated with the second PC (p=2.5E-14).

![Figure 3.33: Graph indicating the contribution of each principal component loading to the variability in the dataset of RA discordant MZ twins.](image-url)
<table>
<thead>
<tr>
<th>PC1</th>
<th>PC2</th>
<th>PC3</th>
<th>PC4</th>
<th>PC5</th>
<th>PC6</th>
<th>PC7</th>
<th>PC8</th>
<th>PC9</th>
<th>PC10</th>
</tr>
</thead>
<tbody>
<tr>
<td>R²</td>
<td>0.21</td>
<td>0.06</td>
<td>0.05</td>
<td>0.02</td>
<td>0.02</td>
<td>0.01</td>
<td>0.01</td>
<td>0.01</td>
<td>0.01</td>
</tr>
<tr>
<td>Cumulative R²</td>
<td>0.21</td>
<td>0.27</td>
<td>0.32</td>
<td>0.34</td>
<td>0.37</td>
<td>0.39</td>
<td>0.40</td>
<td>0.41</td>
<td>0.43</td>
</tr>
</tbody>
</table>

Table 3.22: Contribution of each principal component (PC) to variability of dataset. R² of each principal component indicates the contribution of that individual component while cumulative R² indicates the amount of variability explained by all the PCs up to and including that one.

<table>
<thead>
<tr>
<th>Source of DNA</th>
<th>PC1</th>
<th>PC2</th>
<th>PC3</th>
<th>PC4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sample plate</td>
<td>0.701</td>
<td>0.004813</td>
<td>0.9275</td>
<td>0.5451</td>
</tr>
<tr>
<td>BeadChip</td>
<td>0.9321</td>
<td>0.3589</td>
<td>0.1019</td>
<td>0.2434</td>
</tr>
<tr>
<td>Position on BeadChip</td>
<td>0.08928</td>
<td>0.0003264</td>
<td>0.01795</td>
<td>0.9729</td>
</tr>
<tr>
<td>Gender</td>
<td>0.0332</td>
<td>0.2861</td>
<td>0.08459</td>
<td>0.5868</td>
</tr>
<tr>
<td>Age</td>
<td>0.4439</td>
<td>0.004294</td>
<td>0.8709</td>
<td>0.001399</td>
</tr>
<tr>
<td>Smoking status</td>
<td>0.2887</td>
<td>0.8981</td>
<td>0.8242</td>
<td>0.003478</td>
</tr>
<tr>
<td>Family</td>
<td>0.01251</td>
<td>1.141E-06</td>
<td>0.5883</td>
<td>0.02594</td>
</tr>
<tr>
<td>RA status</td>
<td>0.1739</td>
<td>0.2234</td>
<td>0.3007</td>
<td>0.1929</td>
</tr>
</tbody>
</table>

Table 3.23: P-values for the correlation of the first four PC loadings with possible batch effects and patient characteristics in whole blood RA discordant monozygotic twin methylation analysis.
3.3.2.5. Identification of differentially methylated positions

Differentially methylated positions were identified using a mixed effects model in the CpGassoc package (Barfield, Conneely, and Kilaru 2012). The first and second PCs were included in the model as fixed covariates, while the family ID for each twin pair was included as a random covariate, to adjust for family structure. PC1 and PC2 were used in place of patient characteristics in this analysis because during PC analysis, PC1 did not correlate with any known traits. 13 differentially methylated positions were identified (FDR 5%), one of which (cg23573114) also passed the more stringent Bonferroni significance, as shown in Figure 3.34 and Table 3.24. The most differentially methylated position lies in the N-shelf of a CpG island of the LOC647979 gene, which encodes a miRNA. There was a general trend towards increased methylation in RA affected twins compared to non-RA twins (Table 3.24).
Figure 3.34: Manhattan plot of differentially methylated positions in RA discordant MZ twin pairs in whole blood samples, unadjusted for cell composition. In this epigenome-wide association study, 13 CpG sites were differentially methylated and passed a false discovery rate of 5% (dashed line). One site, cg23573114 passed the more stringent Bonferroni correction for multiple testing (solid line).
Table 3.24: Differentially methylated positions (DMPs) between rheumatoid arthritis discordant monozygotic twins identified in whole blood samples, unadjusted for cell composition.

Table shows the top thirteen differentially methylated positions, the p-value of each site’s association with twin discordance for RA, mean beta-value of methylation in RA and non-RA groups and the chromosomal position of each site, with associated gene annotation. All CpG sites shown were passed false discovery rate of 5%. The top DMP, cg23573114, also passed the Bonferroni significance.

<table>
<thead>
<tr>
<th>DMP</th>
<th>P-value of difference</th>
<th>Mean (SD) β-values in RA twins</th>
<th>Mean (SD) β-values in non-RA twins</th>
<th>Chromosome: physical position (annotation)</th>
</tr>
</thead>
<tbody>
<tr>
<td>cg23573114</td>
<td>3.81E-08</td>
<td>0.69 (0.03)</td>
<td>0.66 (0.03)</td>
<td>Chr20: 34634400 (in gene body of LOC647979)</td>
</tr>
<tr>
<td>cg08520702</td>
<td>1.43E-07</td>
<td>0.90 (0.01)</td>
<td>0.89 (0.01)</td>
<td>Chr8: 142157693 (In gene body of DENND3)</td>
</tr>
<tr>
<td>cg07693617</td>
<td>1.78E-07</td>
<td>0.77 (0.05)</td>
<td>0.74 (0.03)</td>
<td>Chr1: 2081984 (In gene body of PRKCG)</td>
</tr>
<tr>
<td>cg00782811</td>
<td>2.18E-07</td>
<td>0.29 (0.03)</td>
<td>0.27 (0.02)</td>
<td>Chr 6: 46293734 (In TSS of RCAN2)</td>
</tr>
<tr>
<td>cg07636225</td>
<td>2.19E-07</td>
<td>0.86 (0.03)</td>
<td>0.84 (0.02)</td>
<td>Chr 11: 63466987 (In gene body of RTN3)</td>
</tr>
<tr>
<td>cg26246127</td>
<td>3.76E-07</td>
<td>0.06 (0.01)</td>
<td>0.05 (0.01)</td>
<td>Chr 5: 87437400</td>
</tr>
<tr>
<td>cg01901579</td>
<td>4.91E-07</td>
<td>0.86 (0.03)</td>
<td>0.84 (0.03)</td>
<td>Chr 14: 95615731 (In 5’UTR of DICER1)</td>
</tr>
<tr>
<td>cg10652637</td>
<td>5.40E-07</td>
<td>0.87 (0.03)</td>
<td>0.85 (0.03)</td>
<td>Chr 5: 98299078</td>
</tr>
<tr>
<td>cg12332239</td>
<td>6.44E-07</td>
<td>0.57 (0.06)</td>
<td>0.53 (0.05)</td>
<td>Chr 17: 66251609 (In gene body of AMZ2)</td>
</tr>
<tr>
<td>cg17134153</td>
<td>9.20E-07</td>
<td>0.29 (0.06)</td>
<td>0.27 (0.06)</td>
<td>Chr 1: 157670328 (In 5’UTR of FCRL3)</td>
</tr>
<tr>
<td>cg00539261</td>
<td>9.96E-07</td>
<td>0.82 (0.03)</td>
<td>0.81 (0.02)</td>
<td>Chr 1: 27678138 (In gene body of SYTL1)</td>
</tr>
<tr>
<td>cg17149644</td>
<td>1.06E-06</td>
<td>0.07 (0.01)</td>
<td>0.06 (0.01)</td>
<td>Chr 11: 126225276 (in TSS of ST3GAL4)</td>
</tr>
<tr>
<td>cg15703512</td>
<td>1.21E-06</td>
<td>0.07 (0.01)</td>
<td>0.07 (0.01)</td>
<td>Chr 16: 22012565 (in TSS of C16orf65)</td>
</tr>
</tbody>
</table>
3.3.2.6. *Estimation of cell composition*

Cell composition was estimated from the BeadChip methylation data using the Houseman regression calibration algorithm as described previously (section 3.2.1.2.8). The proportion of each cell type in RA-affected twins and non-RA twins is shown below in Table 3.25. Proportions of CD8 T-cells and NK-cells were very similar in both RA and non-RA populations. CD4 T-cells and monocytes showed a trend towards having slightly higher proportions in RA affected twins, though not statistically significant (p>0.05). Proportions of B-cells and granulocytes were significantly different between RA and healthy twins (p<0.05), B-cells were present at higher proportion in RA twins while there was a lower proportion of granulocytes. In RA affected twins CD8 T-cells had the lowest proportion, whereas in healthy twins B-cells were the least common. In both groups, granulocytes formed the largest population, as expected.

<table>
<thead>
<tr>
<th>Cell type</th>
<th>Proportion in RA twins</th>
<th>Proportion in non-RA twins</th>
<th>Difference</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>CD8T</td>
<td>3.55</td>
<td>2.65</td>
<td>0.90</td>
<td>0.22</td>
</tr>
<tr>
<td>CD4T</td>
<td>16.81</td>
<td>12.27</td>
<td>4.54</td>
<td>0.07</td>
</tr>
<tr>
<td>NK cells</td>
<td>4.40</td>
<td>4.55</td>
<td>0.15</td>
<td>0.84</td>
</tr>
<tr>
<td>B cells</td>
<td>4.19</td>
<td>2.24</td>
<td>1.95</td>
<td>0.01</td>
</tr>
<tr>
<td>Monocytes</td>
<td>7.26</td>
<td>6.04</td>
<td>1.22</td>
<td>0.05</td>
</tr>
<tr>
<td>Granulocytes</td>
<td>63.40</td>
<td>71.70</td>
<td>8.30</td>
<td>0.04</td>
</tr>
</tbody>
</table>

Table 3.25: Cell composition estimates calculated from whole blood HumanMethylation450 BeadChip data from RA discordant MZ twins.
3.3.2.7. Identification of differentially methylated positions following adjustment for cell composition differences

The identification of differentially methylated positions analysis (described in section 3.3.2.5) was repeated but with adjustment for cell composition using the cell proportion estimates calculated in section 3.3.2.6. Cell proportions were included as fixed covariates in the mixed effects model along with PC1 and PC2. Family structure was included as a random covariate.

Figure 3.35: Manhattan plot of differentially methylated positions between rheumatoid arthritis discordant monozygotic twins after adjustment for cell composition, identified in whole blood.
<table>
<thead>
<tr>
<th>DMP</th>
<th>P-value of difference</th>
<th>Mean (SD) β-values in RA twins</th>
<th>Mean (SD) β-values in non-RA twins</th>
<th>Chromosome: physical position (annotation)</th>
</tr>
</thead>
<tbody>
<tr>
<td>cg23573114</td>
<td>7.83E-07</td>
<td>0.69 (0.03)</td>
<td>0.66 (0.03)</td>
<td>Chr20: 34634400 (gene body of LOC647979)</td>
</tr>
<tr>
<td>cg05421392</td>
<td>8.41E-07</td>
<td>0.71 (0.07)</td>
<td>0.75 (0.05)</td>
<td>Chr11: 55704601 (TSS of OR5I1)</td>
</tr>
<tr>
<td>cg14783646</td>
<td>1.11E-06</td>
<td>0.92 (0.02)</td>
<td>0.93 (0.02)</td>
<td>Chr7: 4800873 (gene body of FOXK1)</td>
</tr>
<tr>
<td>cg26246127</td>
<td>1.83E-06</td>
<td>0.06 (0.01)</td>
<td>0.05 (0.01)</td>
<td>Chr5: 87437400</td>
</tr>
<tr>
<td>cg20802333</td>
<td>2.65E-06</td>
<td>0.88 (0.03)</td>
<td>0.91 (0.02)</td>
<td>Chr1: 244552382 (3'UTR of C1orf100)</td>
</tr>
<tr>
<td>cg25337513</td>
<td>3.94E-06</td>
<td>0.82 (0.02)</td>
<td>0.83 (0.02)</td>
<td>Chr17: 78800631 (gene body of RPTOR)</td>
</tr>
<tr>
<td>cg03733278</td>
<td>4.97E-06</td>
<td>0.91 (0.02)</td>
<td>0.93 (0.02)</td>
<td>Chr22: 20760922 (gene body of ZNF74)</td>
</tr>
<tr>
<td>cg00782811</td>
<td>5.71E-06</td>
<td>0.29 (0.03)</td>
<td>0.27 (0.02)</td>
<td>Chr6: 46293734 (TSS of RCAN2)</td>
</tr>
<tr>
<td>cg26985354</td>
<td>6.31E-06</td>
<td>0.80 (0.04)</td>
<td>0.83 (0.02)</td>
<td>Chr19: 45567180 (gene body of SFRS16)</td>
</tr>
<tr>
<td>cg01901579</td>
<td>6.78E-06</td>
<td>0.86 (0.03)</td>
<td>0.84 (0.03)</td>
<td>Chr14: 95615731 (5'UTR of DICER1)</td>
</tr>
</tbody>
</table>

Table 3.26: Differentially methylated positions in whole blood samples from RA discordant MZ twins, after adjustment for cell composition. Table shows the top ten differentially methylated positions, the p-value of each sites association with twin discordance for RA, mean beta-value of methylation in RA and non-RA groups and the chromosomal position of each site, with associated gene annotation. None of these sites passed false discovery rate of 5%.
3.3.3. **Validation using rheumatoid arthritis discordant twin PBMC samples**

3.3.3.1. *Quality control*

Quality control was performed on PBMC and whole blood samples together, as described in section 3.3.2.1.

3.3.3.2. *Normalisation*

SWAN normalisation was performed on the samples as described in section 3.2.1.2.2.

3.3.3.3. *Pruning unwanted probes*

Probes that had a detection p-value of more than 0.01 were excluded from analysis (n=11295), along with probes mapping to the X or Y chromosomes (n=11,648). This left 472,924 probes for further analysis. SNP-containing probes were not removed because the comparison groups were composed of MZ twins, which are assumed to have the same SNPs.

3.3.3.4. *Principal components analysis*

PC analysis identified one loading which explained over 20% of the variability in the methylation data and two loadings that explained over 5% of the variability (Figure 3.36). The top four PCs were compared to potential batch effects and patient characteristics (Table 3.27). The first PC correlated strongly with gender (p=1.6E-6). The second PC did not correlate significantly with any of the possible batches or patient characteristics tested. The third PC correlated weakly with position on BeadChip (p=0.03).
Figure 3.36: Contribution of principal components to variability in the HumanMethylation450 dataset of rheumatoid arthritis discordant monozygotic twins in PBMC samples. The horizontal line indicates the threshold for principal component loadings which explain more than 10% of the variability in the dataset.
<table>
<thead>
<tr>
<th></th>
<th>PC1</th>
<th>PC2</th>
<th>PC3</th>
<th>PC4</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Sample plate</strong></td>
<td>0.09606</td>
<td>0.08976</td>
<td>0.6392</td>
<td>0.3943</td>
</tr>
<tr>
<td><strong>BeadChip</strong></td>
<td>0.2759</td>
<td>0.6946</td>
<td>0.06852</td>
<td>0.007917</td>
</tr>
<tr>
<td><strong>Position on BeadChip</strong></td>
<td>0.7772</td>
<td>0.8082</td>
<td>0.03092</td>
<td>0.08854</td>
</tr>
<tr>
<td><strong>Gender</strong></td>
<td>1.615e-06</td>
<td>0.173</td>
<td>0.9028</td>
<td>0.8196</td>
</tr>
<tr>
<td><strong>Age</strong></td>
<td>0.8649</td>
<td>0.4884</td>
<td>0.9835</td>
<td>0.04576</td>
</tr>
<tr>
<td><strong>Smoking status</strong></td>
<td>0.1968</td>
<td>0.3367</td>
<td>0.04552</td>
<td>0.1864</td>
</tr>
<tr>
<td><strong>Family</strong></td>
<td>0.08016</td>
<td>0.09767</td>
<td>0.2962</td>
<td>0.4319</td>
</tr>
<tr>
<td><strong>RA status</strong></td>
<td>0.2359</td>
<td>0.6685</td>
<td>0.8911</td>
<td>0.8646</td>
</tr>
</tbody>
</table>

Table 3.27: P-values of association between the first four principal component loadings and potential batch effects. Gender correlates most strongly with principal component 1 (PC1), while no measured patient characteristics or batch effects correlate with PC2. PC3 is slightly associated with position on BeadChip, while PC4 is weakly associated with BeadChip ID.
3.3.3.5. Identification of differentially methylated positions

Differentially methylated positions were identified in PBMC samples using a mixed effects model in the CpGassoc package. The first three PCs (which each accounted for >5% variability in the dataset) were included as fixed covariates and family ID was included as a random covariate. No CpG sites passed false discovery rate of 5% (see Table 3.28 and Figure 3.37).

Figure 3.37: Manhattan Plot for association between differentially methylated positions and RA status of MZ twins in PBMC samples.
<table>
<thead>
<tr>
<th>DMP</th>
<th>P-value of difference</th>
<th>Mean (SD) β-values in RA twins</th>
<th>Mean (SD) β-values in non-RA twins</th>
<th>Chromosome: physical position (annotation)</th>
</tr>
</thead>
<tbody>
<tr>
<td>cg07380200</td>
<td>2.27E-06</td>
<td>0.86 (0.02)</td>
<td>0.84 (0.02)</td>
<td>Chr15: 48874903 (gene body of FBN1)</td>
</tr>
<tr>
<td>cg11065467</td>
<td>2.56E-06</td>
<td>0.89 (0.02)</td>
<td>0.90 (0.02)</td>
<td>Chr10: 133218425</td>
</tr>
<tr>
<td>cg11938395</td>
<td>6.64E-06</td>
<td>0.90 (0.01)</td>
<td>0.91 (0.01)</td>
<td>Chr12: 123200533 (In first exon of GPR109B)</td>
</tr>
<tr>
<td>cg21077559</td>
<td>1.06E-05</td>
<td>0.02 (0.01)</td>
<td>0.03 (0.01)</td>
<td>Chr4: 122686319 (first exon of TMEM155)</td>
</tr>
<tr>
<td>cg18348142</td>
<td>1.22E-05</td>
<td>0.14 (0.04)</td>
<td>0.16 (0.03)</td>
<td>Chr12: 47474055 (TSS of AMIGO2)</td>
</tr>
<tr>
<td>cg01182873</td>
<td>2.21E-05</td>
<td>0.91 (0.01)</td>
<td>0.89 (0.02)</td>
<td>Chr7: 141464291 (first exon of TAS2R3)</td>
</tr>
<tr>
<td>cg12456927</td>
<td>2.79E-05</td>
<td>0.72 (0.11)</td>
<td>0.75 (0.10)</td>
<td>Chr4: 1522229</td>
</tr>
<tr>
<td>cg05771377</td>
<td>3.18E-05</td>
<td>0.90 (0.01)</td>
<td>0.89 (0.01)</td>
<td>Chr16: 649603 (gene body of RAB40C)</td>
</tr>
<tr>
<td>cg25107608</td>
<td>3.29E-05</td>
<td>0.03 (0.01)</td>
<td>0.04 (0.01)</td>
<td>Chr1: 154155675 (first exon of TPM3)</td>
</tr>
<tr>
<td>cg03396324</td>
<td>3.60E-05</td>
<td>0.92 (0.01)</td>
<td>0.93 (0.01)</td>
<td>Chr3: 184060570 (gene body of FAM131A)</td>
</tr>
</tbody>
</table>

Table 3.28: Differentially methylated positions identified in PBMC samples from RA discordant MZ twins. Table shows the top ten differentially methylated positions, the p-value of each sites association with twin discordance for RA, mean beta-value of methylation in RA and non-RA groups and the chromosomal position of each site, with associated gene annotation. None of these sites passed false discovery rate of 5%. 

252
3.3.3.6. *Estimation of cell composition*

Cell composition of PBMC samples were estimated using the Houseman method in minfi. The proportion of CD8 T-cells, CD4 T-cells, NK-cells, B-cells, monocytes and granulocytes was estimated in RA affected twins and non-RA twins (see Table 3.29). Granulocytes had the lowest proportion in both RA and non-RA groups (4.82 and 6.75 respectively); this is expected as the cell subset is specifically removed during processing to collect PBMCs. CD4 T-cells had the highest proportion (48.16 and 45.26 respectively). The biggest difference between groups was in B-cells (4.15) while the smallest difference was in monocytes. There were no significant differences in cell types between RA and non-RA twins.

<table>
<thead>
<tr>
<th>Cell type</th>
<th>Proportion in RA twins</th>
<th>Proportion in non-RA twins</th>
<th>Difference</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>CD8 T</td>
<td>11.12</td>
<td>12.31</td>
<td>1.19</td>
<td>0.61</td>
</tr>
<tr>
<td>CD4 T</td>
<td>48.16</td>
<td>45.26</td>
<td>2.90</td>
<td>0.55</td>
</tr>
<tr>
<td>NK-cells</td>
<td>12.63</td>
<td>12.29</td>
<td>0.34</td>
<td>0.92</td>
</tr>
<tr>
<td>B-cells</td>
<td>12.67</td>
<td>16.82</td>
<td>4.15</td>
<td>0.19</td>
</tr>
<tr>
<td>Monocytes</td>
<td>8.17</td>
<td>7.81</td>
<td>0.36</td>
<td>0.87</td>
</tr>
<tr>
<td>Granulocytes</td>
<td>6.75</td>
<td>4.82</td>
<td>1.93</td>
<td>0.47</td>
</tr>
</tbody>
</table>

Table 3.29: Cell proportion estimates in RA and non-RA twins.
3.3.3.7. Identification of differentially methylated positions with adjustment for cell composition

Differentially methylated positions were identified using a mixed effects model in the CpGassoc package (Barfield, Conneely, and Kilaru 2012). The first three PCs were included as fixed covariates along with the cell estimates for each sample, as described in section 3.3.2.7. One CpG site (cg00584174) passed false discovery rate of 5% and Bonferroni significance (Figure 3.38 and Table 3.30). Interestingly, the third DMP (cg21984711) lies in the 3’ untranslated region (UTR) of the BCL2 gene, which has previously been found to have increased expression in RA fibroblast-like synoviocytes (Lee et al. 2013).
Figure 3.38: Manhattan plot of differentially methylated positions identified in PBMC samples from RA discordant MZ twins with adjustment for cell composition.
<table>
<thead>
<tr>
<th>DMP</th>
<th>P-value of difference</th>
<th>Mean (SD) β-values in RA twins</th>
<th>Mean (SD) β-values in non-RA twins</th>
<th>Chromosome: physical position (annotation)</th>
</tr>
</thead>
<tbody>
<tr>
<td>cg00584174</td>
<td>5.58E-08</td>
<td>0.85 (0.03)</td>
<td>0.87 (0.03)</td>
<td>Chr11: 63257403 (gene body of HRASLS5)</td>
</tr>
<tr>
<td>cg15509575</td>
<td>1.88E-06</td>
<td>0.90 (0.01)</td>
<td>0.89 (0.01)</td>
<td>Chr4: 2286509 (gene body of ZFYVE28)</td>
</tr>
<tr>
<td>cg21984711</td>
<td>5.15E-06</td>
<td>0.92 (0.02)</td>
<td>0.93 (0.01)</td>
<td>Chr18: 60791714 (3'UTR of BCL2)</td>
</tr>
<tr>
<td>cg00570069</td>
<td>1.72E-05</td>
<td>0.83 (0.01)</td>
<td>0.82 (0.02)</td>
<td>Chr19: 14271335 (gene body of LPHN1)</td>
</tr>
<tr>
<td>cg15649327</td>
<td>1.91E-05</td>
<td>0.91 (0.02)</td>
<td>0.92 (0.02)</td>
<td>Chr1: 59166369 (TSS of MYSM1)</td>
</tr>
<tr>
<td>cg22907789</td>
<td>1.94E-05</td>
<td>0.11 (0.02)</td>
<td>0.12 (0.03)</td>
<td>Chr17: 75567709</td>
</tr>
<tr>
<td>cg24872571</td>
<td>1.94E-05</td>
<td>0.90 (0.01)</td>
<td>0.90 (0.01)</td>
<td>Chr19: 41129881 (gene body of LTBP4)</td>
</tr>
<tr>
<td>cg00758701</td>
<td>3.04E-05</td>
<td>0.81 (0.02)</td>
<td>0.81 (0.03)</td>
<td>Chr13: 49879718</td>
</tr>
<tr>
<td>cg13109045</td>
<td>3.42E-05</td>
<td>0.07 (0.01)</td>
<td>0.06 (0.01)</td>
<td>Chr14: 105218947 (TSS of SIVA1)</td>
</tr>
<tr>
<td>cg08426384</td>
<td>5.07E-05</td>
<td>0.04 (0.01)</td>
<td>0.03 (0.004)</td>
<td>Chr14: 78266734 (5'UTR of ADCK1)</td>
</tr>
</tbody>
</table>

Table 3.30: Differentially methylated positions identified in PBMC samples from RA discordant MZ twins with adjustment for cell composition. Table shows the top ten differentially methylated positions, the p-value of each sites association with twin discordance for RA, mean beta-value of methylation in RA and non-RA groups and the chromosomal position of each site, with associated gene annotation. One of these sites passed false discovery rate of 5% and Bonferroni significance (indicated in bold).
3.3.4. **Summary of results of the investigation into methylation in rheumatoid arthritis discordant monozygotic twins**

This investigation into DNA methylation in RA discordant MZ twins involved epigenome-wide association studies of WB and PBMC samples. In WB samples, 13 CpG sites were differentially methylated with a false discovery rate of 5% without adjustment for cell composition. Of these 13 sites, one also surpassed the more stringent Bonferroni significance. After adjustment for cell composition, no sites passed the FDR threshold. Interestingly, four of the sites identified as significantly differentially methylated in the unadjusted analysis remained in the top ten most differentially methylated positions after adjustment for cell composition, and the top most differentially methylated position was the same whether cell composition was adjusted for or not.

Analysis of the PBMC samples without adjustment for cell composition did not identify any significantly differentially methylated positions. After adjustment for cell composition, one CpG site was differentially methylated and passed the significance thresholds.

The statistically derived cell estimates were notably different between whole blood and PBMC samples, as expected, indicating that this method of estimating cell composition is appropriate. There was no overlap in the top ten DMPs in PBMCs and whole blood for analyses adjusted or unadjusted for cell composition.
3.4. **Investigation into the effect of cryopreservation on DNA methylation in CD3+CD4+ T-cells**

3.4.1. **Patient characteristics**

Samples from 5 healthy volunteers were selected to investigate the effect of cryopreservation of PBMCs on methylation status of CD3+CD4+ T-cells. Volunteers were recruited from the NRHV in Manchester (described in section 2.1.4) after an email request for volunteers was circulated to the NRHV registered participants. One volunteer was female and the ages of participants ranged between 26 and 39 as summarised in Table 3.31.

<table>
<thead>
<tr>
<th>Volunteer ID</th>
<th>Gender</th>
<th>Age</th>
</tr>
</thead>
<tbody>
<tr>
<td>NRHV006</td>
<td>Male</td>
<td>34</td>
</tr>
<tr>
<td>NRHV015</td>
<td>Male</td>
<td>37</td>
</tr>
<tr>
<td>NRHV152</td>
<td>Female</td>
<td>30</td>
</tr>
<tr>
<td>NRHV154</td>
<td>Male</td>
<td>26</td>
</tr>
<tr>
<td>NRHV157</td>
<td>Male</td>
<td>39</td>
</tr>
</tbody>
</table>

Table 3.31: Age and gender of healthy volunteers.
3.4.2. Preparation of CD3+CD4+ T-cells by FACS

Fresh (n=5) and cryopreserved (n=5) samples for FACS analysis were prepared as described in section 2.2.1. FACS was performed at the Faculty of Life Science core facility at the University of Manchester by Mr Michael Jackson. Cells were gated based on their forward and side scatter profiles, an example of which is shown in Figure 3.39. After sorting, samples were again analysed using FACS to assess post-sort purity, an example of which is shown in Figure 3.39. This was performed for the five samples which were freshly prepared, and for the five cryopreserved samples from the same individuals (see section 2.2.1). One sample, from individual ‘NRHV157’ was also subjected to FACS without actually separating out an individual cell population (i.e. all PBMCs were gated for) to assess the effect of FACS sorting itself on DNA methylation. The post-FACS cell counts are shown in Table 3.32 for the PBMC sample which was subjected to FACS without being sorted, and for the samples from five individuals that were sorted with and without cryopreservation of the PBMCs. As expected, there were less viable cells in the cryopreserved samples (shown in Table 3.32 and Figure 3.40). This is likely to be due to cell death during cryopreservation, and also loss of some cells during the cell clean-up stage following cryopreservation (which is required to remove DMSO from cell suspension, as described in section 2.2.1.3). While some cell death is expected following cryopreservation of cells, the levels seen in this study are higher than expected, indicating that these techniques require optimisation prior to use in further studies.
Figure 3.39: Forward and Side scatter plot of PBMC sample before and after sorting, showing gate selection (P2). Plot on the left shows the PBMC sample before sorting, with gate selection for CD3+CD4+ T-cells shown (labelled as P2). The right hand plot shows the post-sort population.

Figure 3.40: Forward and side scatter plots of post-sort CD3+CD4+ T-cells from the same individual, performed on fresh PBMCs (left) and cryopreserved PBMCs (right).
<table>
<thead>
<tr>
<th>Cell type</th>
<th>Individual</th>
<th>Cell count (fresh)</th>
<th>Cell count (cryopreserved)</th>
</tr>
</thead>
<tbody>
<tr>
<td>PBMCs</td>
<td>NRHV157</td>
<td>3,500,000</td>
<td>N/A</td>
</tr>
<tr>
<td>CD3+CD4+ T-cells</td>
<td>NRHV157</td>
<td>1,856,000</td>
<td>132,000</td>
</tr>
<tr>
<td></td>
<td>NRHV152</td>
<td>3,732,000</td>
<td>388,000</td>
</tr>
<tr>
<td></td>
<td>NRHV006</td>
<td>1,490,000</td>
<td>107,000</td>
</tr>
<tr>
<td></td>
<td>NRHV015</td>
<td>1,391,000</td>
<td>220,000</td>
</tr>
<tr>
<td></td>
<td>NRHV154</td>
<td>1,569,000</td>
<td>200,000</td>
</tr>
</tbody>
</table>

Table 3.32: Counts of fresh PBMCs, and CD3+CD4+ T-cell counts from fresh and cryopreserved samples obtained from 10ml blood.
3.4.3. Whole genome DNA methylation profiling of CD3+CD4+ T-cells

3.4.3.1. Quality control

The quality of samples was determined by the distribution of beta values on a kernel density plot (Figure 3.41) which showed that all samples had the expected bimodal distribution. The median log transformed methylated and unmethylated probe intensities were also plotted with a quality threshold of 10.5 (Figure 3.42). There were no obvious outliers in either plot, and no samples were excluded from the analysis. A multiple density scaling (MDS) plot produced using the minfi package showed the distribution of the 1000 most variable positions in each sample. As expected, samples from each individual clustered tightly together (Figure 3.43).

Figure 3.41: Quality control kernel density plot showing distribution of beta values of methylation for each sample in the investigation of the effect of cryopreservation on methylation.
Figure 3.42: Quality control plot of log transformed median methylated and unmethylated signal for each sample in the investigation of the effect of cryopreservation on methylation.
Figure 3.43: Multiple density scaling plot of the distribution of the 1000 most variable positions in samples in the investigation of the effect of cryopreservation on methylation. As expected, samples from each individual cluster tightly together. Samples are colour coded by their sample ID, as shown in the key along the x-axis.
3.4.3.2. **Epigenome-wide association study of CD3+CD4+ T-cells separated from cryopreserved versus fresh PBMCs**

3.4.3.2.1. **Normalisation**

Samples were SWAN normalised, as described in section 3.2.1.2.2.

3.4.3.2.2. **Pruning unwanted probes**

Probes that had a detection p-value of more than 0.01 were considered to have failed and were removed from further analysis. In the CD3+CD4+ T-cell samples, 1424 probes failed detection-p. Probes mapping to the X or Y chromosomes were also removed (n=11,648). This left 473,453 probes to continue into further analysis in the CD3+CD4+ T-cell samples.

3.4.3.2.3. **Principal components analysis**

PCA was performed using the PCAmethods package (Stacklies et al. 2007). Four PCs explained more than 10% of the variability in the dataset each (Figure 3.44). The first four PCs were compared to position on the BeadChip, age, gender, cryopreservation status of sample and individual. Position on array and age correlated weakly with PC2 (p=0.0001) while individual correlated strongly with PC3 and PC4 (8.6E-6 and 1.6E-6 respectively).
Figure 3.44: Principal components and the contribution of each loading to variability in methylation profile dataset of CD3+CD4+ T cells which had been prepared from cryopreserved or fresh samples.

The horizontal line indicates the threshold for principal component loadings which explain more than 10% of the variability in the dataset. Four loadings passed this threshold.

<table>
<thead>
<tr>
<th>Index</th>
<th>R2</th>
<th>Cumulative R2</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>PC1</td>
<td>0.34</td>
</tr>
<tr>
<td></td>
<td>PC2</td>
<td>0.15</td>
</tr>
<tr>
<td></td>
<td>PC3</td>
<td>0.13</td>
</tr>
<tr>
<td></td>
<td>PC4</td>
<td>0.12</td>
</tr>
<tr>
<td></td>
<td>PC5</td>
<td>0.087</td>
</tr>
<tr>
<td></td>
<td>PC6</td>
<td>0.046</td>
</tr>
<tr>
<td></td>
<td>PC7</td>
<td>0.04</td>
</tr>
<tr>
<td></td>
<td>PC8</td>
<td>0.04</td>
</tr>
<tr>
<td></td>
<td>PC9</td>
<td>0.04</td>
</tr>
<tr>
<td></td>
<td>PC10</td>
<td>0</td>
</tr>
</tbody>
</table>

Table 3.33: Contribution of each of the first ten principal components to variability in dataset of cryopreserved and fresh CD3+CD4+ T cells.
<table>
<thead>
<tr>
<th></th>
<th>PC1</th>
<th>PC2</th>
<th>PC3</th>
<th>PC4</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Position on BeadChip</strong></td>
<td>0.443</td>
<td>0.0001009</td>
<td>0.8821</td>
<td>0.9739</td>
</tr>
<tr>
<td><strong>Gender</strong></td>
<td>0.3749</td>
<td>0.194</td>
<td>0.04028</td>
<td>0.2963</td>
</tr>
<tr>
<td><strong>Age</strong></td>
<td>0.8362</td>
<td>0.0001209</td>
<td>0.9507</td>
<td>0.3581</td>
</tr>
<tr>
<td><strong>Cryopreservation</strong></td>
<td>0.05078</td>
<td>0.7135</td>
<td>0.9115</td>
<td>0.9066</td>
</tr>
<tr>
<td><strong>Individual ID</strong></td>
<td>0.2587</td>
<td>0.0007263</td>
<td>8.582E-06</td>
<td>1.576E-06</td>
</tr>
</tbody>
</table>

Table 3.34: Correlation of each of the first four principal components with possible batches and volunteer characteristics in the analysis of cryopreserved and fresh CD3+CD4+ T cells. PC1 was not significantly associated with any batch effects, while PC2 was associated with position on BeadChip, age and individual ID. PC3 and PC4 were correlated strongly with individual ID.
3.4.3.2.4. Identification of differentially methylated positions

Differentially methylated positions were identified using CpGassoc. No probes were differentially methylated with a FDR of 5%, and the most differentially methylated position (cg00281048) had a p-value of 2E-6 (see Figure 3.45 and Table 3.35).

Figure 3.45: Manhattan plot of differentially methylated positions between cryopreserved and fresh samples of CD3+CD4+ T-cells.
<table>
<thead>
<tr>
<th>Rank of CpG</th>
<th>CpG site</th>
<th>p-value</th>
<th>Chromosomal location (gene annotation)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>cg00281048</td>
<td>2.43E-06</td>
<td>Chr.4:80573233</td>
</tr>
<tr>
<td>2</td>
<td>cg12726298</td>
<td>2.84E-06</td>
<td>Chr11:47201474 (In gene body of PACSIN3)</td>
</tr>
<tr>
<td>3</td>
<td>cg21609339</td>
<td>4.13E-06</td>
<td>Chr20:44650270 (In TSS of SLC12A5)</td>
</tr>
<tr>
<td>4</td>
<td>cg14202850</td>
<td>4.23E-06</td>
<td>Chr17:43974869 (In 5'UTR of MAPT)</td>
</tr>
<tr>
<td>5</td>
<td>cg11804928</td>
<td>4.95E-06</td>
<td>Chr17:38220298 (In 5'UTR of THRA)</td>
</tr>
<tr>
<td>6</td>
<td>cg01449136</td>
<td>5.29E-06</td>
<td>Chr14:61109317</td>
</tr>
<tr>
<td>7</td>
<td>cg02251742</td>
<td>6.00E-06</td>
<td>Chr3:156009319 (In gene body of KCNAB1)</td>
</tr>
<tr>
<td>8</td>
<td>cg10971134</td>
<td>6.64E-06</td>
<td>Chr1:56877429</td>
</tr>
<tr>
<td>9</td>
<td>cg05608626</td>
<td>7.14E-06</td>
<td>Chr18:72920877 (in first exon of ZADH2)</td>
</tr>
<tr>
<td>10</td>
<td>cg03596635</td>
<td>8.70E-06</td>
<td>Chr3:127393517 (In gene body of ABTB1)</td>
</tr>
</tbody>
</table>

Table 3.35: Most differentially methylated positions between cryopreserved and fresh samples of CD3+CD4+ T-cells. Table shows the top ten differentially methylated positions, the p-value of each sites association with cryopreservation of cells, and the chromosomal position of each site, with associated gene annotation. None of these sites passed false discovery rate of 5%.
3.4.3.2.1. Correlation of global DNA methylation levels

The SWAN normalised pruned datasets were used to calculate the correlation of each pair of samples (cryopreserved versus fresh from each individual). The Pearson’s correlation for each pair is shown below in Table 3.36. All sample pairs correlated extremely well with Pearson's correlation values of 0.99 and p-values of <2.2E-16. An example of the correlation of methylation values for individual NRHV006 is shown by the cigar plot in Figure 3.46.

<table>
<thead>
<tr>
<th>Individual</th>
<th>Pearson's correlation</th>
</tr>
</thead>
<tbody>
<tr>
<td>NRHV154</td>
<td>0.996</td>
</tr>
<tr>
<td>NRHV152</td>
<td>0.994</td>
</tr>
<tr>
<td>NRHV015</td>
<td>0.996</td>
</tr>
<tr>
<td>NRHV006</td>
<td>0.996</td>
</tr>
<tr>
<td>NRHV157</td>
<td>0.996</td>
</tr>
</tbody>
</table>

Table 3.36: Pearson's correlation for methylation values in cryopreserved versus fresh samples from each individual in the investigation of effect of cryopreservation on methylation.
Figure 3.46: Cigar plot of correlation of methylation values in cryopreserved (y-axis) versus fresh (x-axis) samples from one individual (NRHV006)
3.4.3.3. **Epigenome-wide association study of PBMCs exposed to FACS versus not exposed to FACS**

3.4.3.3.1. **Normalisation**

Samples were SWAN normalised as described in section 3.2.1.2.2.

3.4.3.3.2. **Pruning unwanted probes**

Probes with a detection p-value of >0.01 were considered to have failed and were removed from analysis (n=263). Probes mapping to the gender chromosomes were also removed (n=11,648). This left 473,682 probes for further analysis in the PBMC samples.

3.4.3.3.3. **Correlation of global DNA methylation levels**

The effects of the FACS process on global DNA methylation levels were assessed by correlating the DNA methylation levels of all probes in the two samples. They correlated extremely well with a Pearson’s correlation coefficient of >0.99 (p<2.2E-16). A cigar plot of the correlation is shown below in Figure 3.47.
Figure 3.47: Cigar plot of methylation values in PBMC sample passed through FACS (x-axis) versus a PBMC sample from the same individual that wasn’t exposed to FACS (y-axis).
3.4.4. **Summary of cryopreservation experiment results**

In summary, I have found that cryopreservation of PBMC samples does not significantly alter DNA methylation in CD3+CD4+ T-cells. Global correlation of samples was very high, and while there were small differences identified at individual sites, none of these exceeded the significance threshold. This shows that cryopreservation of PBMC samples for future use in epigenetic studies is a process that will not significantly alter the outcome of the studies. I have also found that exposure of cells to the high level of pressure involved in FACS does not alter the DNA methylation significantly. While individual sites could not be interrogated, global correlation of methylation was >0.99. Any slight differences in methylation between FACS versus non-FACS samples may be due to the removal of dead cells during FACS. These findings could impact the design of future epigenetic studies as they confirm cryopreservation of cells is a technique appropriate for use in investigation of DNA methylation.
4. Discussion
4.1. Overview of findings

This thesis describes three investigations; DNA methylation as a biomarker of response to the anti-TNF therapies etanercept and adalimumab, differences in DNA methylation between RA discordant MZ twins, and a study measuring the effect of cryopreservation of cells on methylation in CD3+CD4+ T cells.

In the first investigation of DNA methylation in whole blood as a predictor of anti-TNF response in patients with RA, I found four CpG sites that were significantly differentially methylated between responders and non-responders to etanercept, while none were significantly different in DNA samples from patients with extreme response phenotypes to adalimumab. Of particular interest, a region of the LRPAP1 gene, containing five differentially methylated positions, was the most differentially methylated locus in the study of etanercept response. When data from the etanercept and adalimumab studies were combined in order to identify predictors of response to anti-TNF drugs as a class, no individual sites showed statistically significant evidence for differential methylation; however, at suggestive levels of association, a region overlapping the CRYZ/TYW3 genes was differentially methylated. The technical validation experiments using pyrosequencing in both studies showed that the array measurements of methylation were reliable but independent replication of the findings in separate groups of responders and non-responders to both drugs will be required to confirm these associations. Nonetheless, this is the first study to identify candidate DNA methylation sites as predictors of response to anti-TNF therapies.

In the largest investigation of DNA methylation differences in disease discordant twins to date, 13 CpGs were differentially methylated in whole blood; however these were no longer significant after adjustment for cell composition. In contrast, PBMC samples from disease discordant twins identified no significant differences in DNA methylation until cell composition was corrected for.
In the study investigating the effect of cryopreservation on methylation, I found that it did not have a substantial effect, showing it is a suitable technique for epigenetic studies. The finding is important as it removes a potential barrier for future studies using cell subsets as cryopreservation will be necessary prior to cell separation in many instances in order that experiments can be performed in batches.

4.2. **Strengths and weaknesses of studies**

4.2.1. **Establishing techniques in the lab**

The studies described in this thesis are the first investigations of DNA methylation to be performed in our group. Because of this, prior to performing the main investigations, it was important to establish the techniques, and ensure their reliability and reproducibility. To do this, I performed bisulphite conversion on several samples in duplicate, to provide biological and technical replicates to measure bisulphite conversion efficiency and the reliability of pyrosequencing as a method of measuring methylation. I validated the pyrosequencing measurements using Sanger sequencing and found that pyrosequencing was more accurate and quantitative, but the results correlated well. Bisulphite conversion and pyrosequencing were found to be reliable and reproducible between both biological and technical replicates, indicating the techniques could be used in subsequent EWAS studies and validation. One assay showed an unexpected large peak at the beginning of the pyrosequencing run, which if encountered in future could be further investigated by sequencing the non-bisulphite converted DNA in the region to ensure the reference sequence (from which the assays were designed) is correct. The stringent quality control steps to ensure that findings generated on one platform could be validated on another, was a strength of the study and add confidence that the findings are robust.
4.2.2. Methylation in anti-TNF response

4.2.2.1. Patient selection and study design

The treatment response studies described in this thesis relied on samples from patients recruited to the BRAGGSS longitudinal cohort. This prospective cohort provides a unique resource for the investigation of treatment response and is a major strength of my study for several reasons. First, this is the largest such cohort world-wide meaning that patients could be matched wherever possible to reduce the effect of baseline confounders that could have affected treatment response, such as gender. It also meant that analysis could be restricted to the extreme response phenotypes, thereby enhancing the power of the study to detect differences. Second, patients are recruited to the study before being treated with a biologic drug and they are then followed prospectively over a 12 month period to assess outcome. The longitudinal nature of the study allows causality to be inferred as all the DNA methylation testing was performed on pre-treatment samples whilst response to therapy was assessed at 3 months. Third, a large amount of clinical data is also collected from patients, again improving the opportunities to match baseline clinical features and to correct for known confounders such as baseline disability (HAQ) score. Finally, the patients are followed up at set time-points using standardised and recognised measures of treatment response such as the change in DAS28 or EULAR response criteria.

The two treatment response studies described in the thesis are pilot studies, intended to identify DNA methylation changes which have effect sizes large enough to be identified in a relatively small number of patients and, therefore, likely to be of clinical importance. The studies are relatively large in the context of methylome-wide investigations of autoimmune disorders, and have more than 90% power to detect changes in methylation of 10% (assuming 12% or less variance). However, a review by Rakyan et al which discussed study design of EWAS suggested that in
order to reliably detect alterations in DNA methylation using the HumanMethylation450 BeadChip, assuming a methylation odds ratio of 1.5, a sample size of 400 cases and 400 controls is required, and to achieve good power, 800 of each is preferable (Rakyan et al. 2011b). Despite that, epigenetic studies with much smaller sample sizes have identified key differentially methylated genes; for example a study with just 65 heavy smokers and 56 non-smokers identified F2RL3 methylation as being related to smoking (Breitling et al. 2011), which has since been validated in a study of 3,588 individuals (Zhang et al. 2014). The sample size needed to have adequate power to detect large, arguably more clinically important, effects are lower; given that I aimed to identify differentially methylated positions that could translate to a clinical test to inform targeting of therapies, it could be argued that the sample sizes used in the current study are sufficient. Furthermore, a strength of this study is that patients with extreme response phenotypes were selected for investigation, increasing the power to identify differences in DNA methylation between good and poor responders. A disadvantage of that approach is that it may reduce the applicability of the results to the wider population, as intermediate responders were not included in the analysis. It is not possible to determine the sensitivity and specificity of the loci identified to predict response using this study design as all patients would need to be included in such an analysis. However, the approach of testing extreme phenotypes also has the advantage of reducing the chance of misclassification of response. Misclassification of response is more common in patients who are intermediate responders, particularly because DAS28 is a composite score including both objective and subjective measures of response. Previous work has shown that psychological factors, such as anxiety and depression, show greater correlation with the subjective components of DAS28 and this makes interpretation of response in intermediate responders difficult; by way of example, synovial inflammation may respond to therapy but the patients may be recorded as being intermediate
responders because of a high tender joint count and VAS score due to concomitant depression (Cordingley et al. 2014).

A further factor that had to be considered in the study design is the expense of the BeadChips used for the EWAS, which limited the sample numbers that could be tested. By selecting extreme response phenotypes, the power of the study was maximised whilst keeping costs within reasonable limits. As EWAS technology reduces in price, inevitably it will become feasible to study larger sample sizes, allowing identification of variants with smaller effects.

A disadvantage of using extreme response phenotypes in the selection criteria is that the pool of patients from which the study participants could be selected was reduced. Ideally, the patients in the response groups would have been matched for characteristics which can alter DNA methylation and have previously been associated with treatment response, such as age, gender, baseline DAS28 score and treatment with concurrent DMARDs (Ellis et al. 2012; Hyrich et al. 2006; Kristensen et al. 2008; Nihal, Wu, and Wood 2014) as well as ACPA status and duration of symptoms. Whilst this was considered where possible in the selection of patients for the etanercept study, there were fewer patients in the extreme response categories for the adalimumab EWAS. This resulted in selecting patients with a significant difference in treatment with concurrent DMARDs between responders and non-responders at baseline (see section 3.2.2.1). To reduce the impact of not being able to match for age, gender, baseline DAS28 score and concurrent DMARD treatment, gender specific probes were removed from the analysis and the other characteristics were adjusted for in the mixed effects model during the identification of DMPs.

The anti-TNF response studies described were performed using whole blood DNA samples, which could be viewed as a disadvantage as differences in DNA
methylation between cell types is recognised (Reinius et al. 2012). Hence, differences in DNA methylation occurring in smaller cell populations may be masked and more difficult to identify. However, I specifically chose to use whole blood as the ultimate aim is to identify a DNA methylation signature which can be used as a clinical predictor of response to anti-TNF drugs, which could be developed into a simple blood test. It is much more appropriate for a clinical test to be performed using whole blood, as the cost and practicality of performing a cell-specific test is unfeasible in a clinical setting.

Many recent studies have highlighted the importance of accounting for cell composition during the analysis of DNA methylation in heterogeneous samples such as whole blood (Jaffe and Irizarry 2014; Liu et al. 2013; Reinius et al. 2012). In order to assess the effect of cell composition on my dataset, cell populations were estimated using a statistical approach that uses the methylation signature of separated cells from a reference dataset to infer cell composition (Houseman et al. 2012; Reinius et al. 2012). The cell estimates for responders and non-responders to the anti-TNF drugs were not statistically different, indicating that cell composition should not affect the results. To confirm this, I performed the analysis with adjustment for cell composition by including the cell composition estimates for each sample in the mixed effects linear regression model during identification of differentially methylated positions. In the etanercept study, 14 of the top 20 DMPs remained the same after adjustment for cell composition. The p-values of the association changed slightly, with the p-value of the top DMP becoming slightly less significant, while generally the p-values of other probes were slightly more significant. After adjustment for cell composition there was an increase in the number of probes passing an FDR of 5%. This is in contrast to the findings of a notable EWAS of RA in cases versus healthy controls, in which the significance of associations was greatly reduced after adjustment for cell composition (Liu et al.}
The difference in effect could be due to a much greater impact of cell composition on EWAS results in that study, as RA pathogenesis is known to be influenced by T and B cells and it is not surprising that differences in cell composition were observed between cases with disease and healthy controls without (Smolen and Steiner 2003). An advantage of my study design is that cell composition is likely to be having less of an effect on my dataset, as all participants have active RA so are likely to have similar cell compositions, as supported by the results of my cell composition estimations. In the adalimumab EWAS the results with and without adjustment for cell composition correlated less well, with just 5 DMPs overlapping in the two analyses. However the top DMP which maps to the *PDZD8* gene, remains the most significant site in both analyses. This suggests that cell composition could have been having more of an influence on the results of the adalimumab EWAS than the etanercept EWAS, possibly due to the smaller number of non-responders in the adalimumab study, or due to the difference in treatment with concurrent DMARDs between groups. With regards to p-values, the opposite effect was seen in the adalimumab analysis as the top DMP increased in significance after adjustment for cell composition (surpassing Bonferroni correction), whereas generally the DMPs had reduced significance.

Some authors have argued that biomarkers of response are more likely to be identified in the target tissue, in this case, the synovium. However, a synovial biopsy is invasive, is associated with risks and cannot be performed in all patients (particularly as RA is relapsing and remitting in nature so synovitis may not be present when the patient attends clinic). A blood test would be far simpler to perform and testing of whole blood would be easier to standardise across different laboratories, making it the ideal tissue to study for response-prediction biomarkers.
4.2.2.2. **Pyrosequencing validation**

The findings from the two anti-TNF response EWAS studies were validated by the use of a different technical approach for measurement of methylation; pyrosequencing. The close correlation of results for the two techniques increased the confidence in findings of the EWAS. Validation was of particular importance as this was the first time these technologies had been used in our lab for methylation detection, so validating both pyrosequencing and the BeadChip technologies was important to ensure the reproducibility of methylation measurements. Another advantage of performing the validation was to measure methylation on a platform which is less sensitive to confounding factors such as batch effects, a recognised problem in measurement of methylation on the BeadChip arrays. A challenge in the validation study was due to the fact that some of the loci selected for validation had small mean changes in DNA methylation within the dataset (eg cg10165222, the top DMP in adalimumab response, had a mean methylation difference of just 2% between responders and non-responders), and pyrosequencing is not sensitive enough to reliably detect such small differences. Previous studies have determined that pyrosequencing should not be used to detect differences in methylation less than 5% (Dejeux et al. 2007; Tost, Dunker, and Gut 2003). Since the majority of sites tested had very subtle differences in DNA methylation between response phenotypes, especially in the adalimumab EWAS, pyrosequencing may not have been a sensitive enough platform for detecting small changes. This explains why the correlation of pyrosequencing and BeadChip results was poor at some sites when measured individually, particularly in the adalimumab study. However, when the results for all adalimumab validation sites were analysed together (ie, pyrosequencing and BeadChip results for all 5 assays compared simultaneously), the correlation was much better ($R^2=0.91$), as the combined assays represented a wider distribution of methylation values.
4.2.3. Methylation in rheumatoid arthritis discordant monozygotic twins

4.2.3.1. Patient selection and sample type

Epigenetic studies of disease performed in cases and unrelated controls are confounded by underlying differences in genetic sequence, which is known to affect DNA methylation (Liu et al. 2013). The greatest strength of my study is that it involves RA discordant MZ twins, allowing investigation of DNA methylation in RA cases with healthy individuals perfectly matched by age, gender, genetic variation and many environmental factors. This should allow the identification of methylation differences associated with disease that are not mediated by genetic sequence (Liu et al. 2013). While such methylation differences could be causative of disease or due to disease pathogenesis, they could also be influenced by treatment. While the study allows a cross-sectional insight into DNA methylation in RA, a limitation of the design is that longitudinal data has not been collected on the twins. Therefore, while at the time of sample collection the twins were discordant for RA, there is no data to show whether or not the healthy twins have developed RA in the future. This could be a limitation of the study if RA methylation changes precede symptomatic disease and diagnosis, as has been found in a study of type 1 diabetes (Rakyan et al. 2011a). As such, the findings of this study should be considered preliminary, and the limitations of the cross-sectional study design, based on diagnosis of RA at one point in time, should be recognised.

The samples investigated in this study were obtained from whole blood (n=62 pairs) and PBMCs that had been cryopreserved (n=17 pairs). A strength of the design is that it allowed investigation of methylation in two different types of samples and allowed identification of methylation differences in different cell types, which may not have been identified in whole blood alone; however a weakness is that the datasets were so different due to different cell composition that they had to be analysed separately. While this reduced power in the main analysis, it also provided a
potential validation cohort for differences that were detected in either study. It should be noted, however, that the power of the PBMC analysis was limited due to the small sample size and, in fact, none of the methylation differences observed in the whole blood analyses were replicated in the PBMC study. DNA from both whole blood and PBMCs contain mixed cell populations, and while this can make it difficult to identify methylation differences which are present in small cell populations, it also provides an insight into the large changes which can be detected in a heterogeneous sample. However, as recent studies have found disease-associated methylation can be tissue specific, the study design limited the ability to investigate methylation differences in cell-subsets (Glossop et al. 2014; Grundberg et al. 2013).

Nonetheless, several previous investigations into disease discordant MZ twins have investigated DNA methylation using whole blood or PBMC DNA in smaller sample sizes (Dempster et al. 2011; Javierre et al. 2010; Selmi et al. 2014; Wong et al. 2014). Many existing collections consist of whole blood DNA samples, and the slow rate at which samples are collected adds a practical difficulty to the collection of cohorts with DNA from individual cell types. Such cohorts would be ideal for epigenetic studies of disease and would allow identification of subtle disease-associated differences between twin pairs; however due to the practical issues, such sample collections are small and rare. The disadvantage of studies which investigate individual cell types is that if the most important cell type in disease pathology is unknown, important methylation differences could be missed by not choosing the appropriate cell type for the study. An advantage of my study design is that it takes all DNA-containing blood cells into account.
4.2.3.2. Study design

My study involved 79 RA discordant MZ twin pairs which is relatively large in the context of epigenetic studies, and is much larger than previous epigenetic twin studies of disease discordant MZ twins (Dempster et al. 2011; Dempster et al. 2014; Gervin et al. 2012; Javierre et al. 2010; Rakyan et al. 2011a; Selmi et al. 2014). Sample size is a fundamental issue in studies of DNA methylation in disease discordant twins, due to the rarity of such twin pairs and, hence, the sparsity of appropriate sample collections. The study would be strengthened by inclusion of twin pairs who were both healthy, to measure the amount of variation in methylation expected within healthy twin pairs. It would also be interesting to include both MZ and DZ twin pairs discordant for disease, to allow estimation of the genetic contribution to methylation variation (Gervin et al. 2011).

In methylation analysis of MZ twins, the within-pair differences identified can be very subtle (Gervin et al. 2011). In order to conserve within-pair differences as much as possible, twin pairs were positioned next to each other on the arrays. An advantage of this approach is that this reduced the batch effects of BeadChip. Previous studies of DNA methylation in twins (in both BeadChip analysis and bisulphite sequencing) have employed the same approach as I did, matching twin pairs within batches (Dempster et al. 2011; Dempster et al. 2014; Gervin et al. 2011) while other studies have chosen to randomly allocate samples across batches to reduce the inflation of the effect of shared environment (Grundberg et al. 2013), however this is likely to increase the identification of false positives due to batch effects. Some large studies which randomly allocate samples have encountered the problem that ComBat adjustment for batch effects can only be used if there is more than one sample in each batch, i.e. if only one sample in a study is on a particular BeadChip or position on BeadChip, ComBat cannot be employed to adjust that batch.
4.2.4. Investigation into effect of cryopreservation on methylation

The current study is the first investigation into the effect of cryopreservation of cells on the DNA methylome. Due to the importance of investigating individual cell types to detect methylation variants which occur in individual cell populations, it is increasingly common that sample collections are subject to cryopreservation of cells such as PBMCs for later separation and use in epigenetic studies. It is already known that environmental factors affect DNA methylation in cells, including lab techniques such as long-term cell culture (Ziller et al. 2013). Despite this, no studies have investigated whether cryopreservation of PBMCs alters the methylome.

In this study I showed that there is no significant alteration in methylation at any individual sites, and that the correlation of methylation in fresh versus cryopreserved cells is extremely good. The main weakness of my study is the small sample size, with only five individuals included. A strength of this study is that methylation was investigated in an individual cell type (CD3+CD4+ T cells). Due to the variation of methylation in different cell types (Reinius et al. 2012), this gives more power to detect changes in methylation due to cryopreservation. This also represents a weakness, as the results only reflect the effect of cryopreservation on a specific cell type, which may be different in other cell types. There was a large loss of viability of cells following cryopreservation of PBMCs, which reduced the number of cells available for methylation analysis. The proportion of cells which had a loss of viability was higher than expected, indicating that the protocol used for isolation and cryopreservation of cells should be optimised prior to its implementation in cohort collections of PBMCs. In the current study, loss of viability should not affect the results as the main consequence was a reduced quantity of DNA, however DNA concentration was normalised across all samples prior to bisulphite conversion, removing this confounder.
Both samples (cryopreserved versus fresh) were taken from each individual on the same day, in the same sitting. This is an advantage as it controls for variables which could affect methylation, such as diet and a plethora of environmental differences. A weakness in the study design is that there are no technical replicates present in the study, so it is difficult to ascertain how much variation would occur between samples simply due to differences in array intensity or bisulphite conversion efficiency. In the future it would be interesting to measure the effect of other lab techniques on the DNA methylation of cells, such as exposure to different temperatures and chemicals. It would also be interesting to characterise the sensitivity of individual cell types to environment, as it is likely that short lived cells are less adaptable to environment.
4.3. **Meaning of studies: possible mechanisms and implications**

4.3.1. **Anti-TNF response studies**

4.3.1.1. *Etanercept*

In the study of treatment response to etanercept, four CpG sites were found to be differentially methylated with an FDR of 0.05. Interestingly, in all variations of the analyses of these studies, the most significantly differentially methylated site remained the same. This DMP (cg04857395) maps to a CpG island in the low density lipoprotein receptor-related protein associated protein 1 (*LRPAP1*) gene body. This site was hypermethylated in non-responders compared to responders to etanercept by 9%, which in the context of complex diseases is a relatively large difference (Rakyan *et al.* 2011a). This CpG island also contains another DMP in the N-shore, and when SNP-containing probes were added into the analysis, another three sites in the same CpG island were found to be differentially methylated. All five of these DMPs map within a 693bp region making this a DMR (Rakyan *et al.* 2011b). Hypermethylation in the gene body is typically associated with increased gene expression (Ball *et al.* 2009). *LRPAP1* encodes a protein which is a chaperone of low density lipoprotein receptor-related protein 1 (*LRP1*) (Willnow *et al.* 1995), and LRPAP1 deficiencies have been found to reduce levels of LRP1, due to increased LRP1 degradation in the endoplasmic reticulum (Aldahmesh *et al.* 2013). LRP1 is known to have a role in inflammation by regulating monocyte recruitment and expression of regulatory cytokines including TNF-α (Arandjelovic *et al.* 2007; Gaultier *et al.* 2008; Gorovoy *et al.* 2010; LaMarre *et al.* 1991; Staudt *et al.* 2013; Wu, Patel, and Pizzo 1998). One study found that LRP1’s anti-inflammatory activity is due to the downregulation of TNFR-1, and found that a TNF-α neutralising antibody inhibited the activity of NF-κB, selectively in LRP1-deficient cells (Gaultier *et al.* 2008). This suggests that in good responders to etanercept, reduced DNA methylation in the *LRPAP1* gene body may lead to reduced expression of LRPAP1,
in turn resulting in reduced levels of LRP1, allowing improved binding of TNF and leading to good response to etanercept (proposed mechanism shown in Figure 4.1). This proposed mechanism is supported by a study which found increased cell surface expression of LRP1 (also known as CD91) in CD3-positive T-cells in non-responders to the anti-TNF infliximab after 6 weeks of treatment (Eriksson, Rantapaa-Dahlqvist, and Sundqvist 2010). The effect was not seen in monocytes, suggesting that infliximab response is driven by CD3 T-cells. The mechanism by which LRP1 could inhibit etanercept binding is unknown, but it could potentially act as a physical blockade preventing the interaction between etanercept and TNF (Figure 4.1).
Figure 4.1: Proposed mechanism by which differential DNA methylation in the LRPAP1 gene body may affect response to etanercept in patients with rheumatoid arthritis. This study has identified higher gene body methylation of the LRPAP1 gene in non-responders to etanercept. I propose that the increased methylation leads to increased expression of the gene, and a subsequent increase in LRP1 protein which is chaperoned by LRPAP1. The increased LRP1 protein could be inhibiting action of etanercept, potentially through physical blockade of the binding site preventing TNF binding.
The 14th most differentially methylated position (cg20066782, \( p=3.13 \times 10^{-6} \)) was also identified as being differentially methylated (\( p=5.57 \times 10^{-8} \)) in the largest EWAS of RA cases versus healthy controls, discussed in section 1.4.7.2 (Liu et al. 2013). Interestingly, this site maps to within the second exon of the gene body of *LRRC17*, a gene previously associated with the NFκB pathway in osteoclast differentiation (Kim et al. 2009). Another DMP identified in the etanercept analysis maps to the *NOTCH4* gene, expression of which has previously been found to be regulated by TNFα in endothelial cells during inflammation (Quillard et al. 2010). The twentieth DMP (cg27400447) in this study maps to the citrate lyase beta like (CLYBL) gene, which was previously associated with ACPA-negative RA in a GWAS study of RA cases versus healthy controls (Bossini-Castillo et al. 2014).

Several of the top most differentially methylated positions map to non-coding intergenic regions. While the role and importance of intergenic methylation is not yet clear, it appears to be involved in repressing the expression of potentially harmful transposable and viral elements (Schulz, Steinhoff, and Florl 2006; Walsh, Chaillet, and Bestor 1998). Extreme cellular conditions have been found to influence intergenic DNA methylation (Ziller et al. 2013), so it is possible that these differences in DNA methylation are due to the cellular environment experienced in RA, which may differ between individuals with the disorder. Long-range epigenetic silencing which covers multiple genes has been identified in cancer (Coolen et al. 2010; Frigola et al. 2006), and is a possible mechanism by which these intergenic methylation differences may control expression of genes over a large distance, however further analysis and functional studies would be necessary to confirm this.
4.3.1.2. **Adalimumab**

In the adalimumab response EWAS, no sites passed the FDR of 5% when uncorrected for cell composition; however, after correction, the top DMP (cg10165222, which is the top DMP in both analyses) surpassed the threshold. This site maps to the **PDZD8** gene, which has been linked to cytoskeletal modelling and retroviral infection (Guth and Sodroski 2014; Henning et al. 2010; Henning et al. 2011). Interestingly, another PDZ domain gene, **PDZD2**, has previously been associated with response to anti-TNF response in a GWAS involving 566 RA patients (Plant et al. 2011).

Interestingly, the second, fourth and eighth most differentially methylated sites identified (cg09121920, cg05857826 and cg15188808 respectively) were previously found to be differentially methylated between RA cases and healthy controls (Liu et al. 2013). This indicates that RA susceptibility DMPs could also be playing a role in treatment response. Two of these persist in the top 10 DMPs after adjustment for cell composition. Cg09121920 maps to the **AXIN1** gene, which encodes a protein involved in Wnt signalling and has been associated with control of inflammatory response during infection (Zhang et al. 2012). Cg05857826 maps to a locus overlapping the transcription start site of ‘protein O-fucosyltransferase 1’ (**POFUT1**) and an intron in the gene body of ‘pleomorphic adenoma gene-like 2’ (**PLAGL2**). Interestingly, **PLAGL2** encodes a zinc-finger transcription factor which is also involved in Wnt signalling (Zheng et al. 2010), and has been found to regulate expression of **NCF2** (which encodes the nicotinamide adenine dinucleotide phosphate (NADPH) oxidase cytosolic component p67phox) in response to TNF-α. This consequently controls p67phox and thus NADPH oxidase activity (Ammons et al. 2007). **POFUT1** is involved in Notch signalling (nnani-Akollor et al. 2014; Okamura and Saga 2008; Stahl et al. 2008).
The 19th DMP, cg26078407, maps to the histone deacetylase 4 (HDAC4) gene. The HDAC4 protein, which is encoded by this gene, has previously been found to accumulate on the promoter of matrix metalloproteinase-1 (MMP-1) in RA synovial fibroblasts following downregulation of MMP-1 (Maciejewska-Rodrigues et al. 2010). The downregulation of this gene by sentrin-specific protease 1 (SENP1) in RA synovial fibroblasts was associated with reduced invasiveness. In cells which were silenced for HDAC4, SENP1 did not control expression of MMP-1, indicating HDAC4 was vital for MMP-1 regulation. The CpG site associated with cg26078407 maps to the gene body of HDAC4, so an increased level of methylation in non-responders to adalimumab would be expected to upregulate the gene. Previous studies have found that HDAC4 is integral in the repression of MMP-1, so it is possible that the change in DNA methylation causes MMP-1 to be repressed in non-responders to adalimumab. This could be affecting the invasive phenotype of synovial fibroblasts in RA, which could consequently be affecting response to adalimumab by increasing inflammation (Maciejewska-Rodrigues et al. 2010). This supports previous findings in animal models and human RA synovial fibroblasts that HDAC inhibitors could be used to reduce synovial inflammation and cartilage destruction in RA (Joosten et al. 2011; Jungel et al. 2006; Lin et al. 2007; Manabe et al. 2008; Nakamura et al. 2008; Nishida et al. 2004). HDAC inhibitors are already in use in the treatment of cutaneous T cell lymphoma and in a trial for treatment of juvenile idiopathic arthritis (JIA), were found to have significant therapeutic benefits (Duvic and Vu 2007; Vojinovic and Damjanov 2011).
4.3.1.3. Combined etanercept and adalimumab

The combined analysis of the etanercept and adalimumab EWAS data was performed to identify any shared indicators of response to anti-TNFs. In one sense, the analysis had much greater power to detect DMPs than the individual studies due to the larger number of samples; however this was hampered by the strong batch effect between studies. The etanercept and adalimumab studies were performed a year apart and this batch effect is apparent in the data, as identified during PCA. The batch is also correlated with the therapy investigated, so the first principal component correlates with both factors strongly (p<10^{-16}). The study population is also less homogenous since it includes individuals who responded well/poorly to different drugs, which while they both act on the TNF pathway, have different mechanisms and so should be investigated separately. One potential way of combating this batch effect and identifying methylation sites which are involved in treatment response in both therapies, but not necessarily in the same direction, is to perform a meta-analysis of the two datasets.

No CpG sites passed a false discovery rate of 5%, and indeed all had very weak statistical evidence for association (p>10^{-5}). The p-values could have been deflated due to adjustment of such a large batch effect using ComBat analysis. Although no sites were significant, a 700bp region which contained 8 DMPs was identified in a region overlapping the CRYZ and TYW3 genes. Interestingly this region was associated with circulating resistin levels, a hormone which has previously been linked with inflammation and type 2 diabetes risk (Qi et al. 2012). More recently, a study has identified the region as being differentially methylated in RA discordant MZ twins (Svendsen et al. 2014). An RA EWAS also identified a CpG (cg09142399) in the region as differentially methylated (p=1.59E-11) between RA cases and unrelated controls (Liu et al. 2013). While the studies support the findings of my work, due to the confounding of batch and different drugs, this result should be
treated with caution until larger EWAS studies of anti-TNF response with fewer batch effects can be performed.
4.3.2. Rheumatoid arthritis discordant monozygotic twin study

Estimation of cell composition in the RA and healthy twins using the Houseman 2012 method indicated all cells in both sample types were within the expected ranges. In both whole blood and PBMC samples, there was a higher proportion of CD4+ T cells in RA twins, which have been suggested to be the cell type driving RA susceptibility due to cell specific histone marks overlapping RA-associated SNPs (Trynka et al. 2013). Furthermore, another study very recently found that differential methylation causes compromised function of Treg cells in RA (Cribbs et al. 2014). This supports the Houseman method as an appropriate technique to produce accurate measurements of cell composition from methylation data.

The first principal component accounted for a large proportion of variability in the dataset, however did not correlate with any batch or patient characteristics measured. Two possible sources of this variability are disease duration and autoantibody status, so in future it would be interesting to investigate if these characteristics correlate with the first principal component.

The DMPs identified in this study had very small differences in methylation between the RA and healthy twin groups. In methylation studies of cancer, methylation differences are usually very large and it is generally accepted that a difference of 10% is required for a DMP to be considered significant. By contrast, non-cancerous complex disorders do not involve such large changes to the genome and epigenome, and many studies have measured small differences in methylation in complex disorders including type 2 diabetes and systemic lupus erythematosus (Bell et al. 2010b; Bell et al. 2010a; Feinberg et al. 2010; Fraga et al. 2005; Javierre et al. 2010; Kaminsky et al. 2009; Nguyen et al. 2010; Rakyan et al. 2011a). The largest investigation to date of DNA methylation in RA (which involved 354 cases and 337 controls) also reported very small differences of around 1-6% between cases and unrelated healthy controls (Liu et al. 2013).
Previous twin studies investigating genome-wide methylation using the HumanMethylation27 BeadChip (the predecessor of the HumanMethylation450 BeadChip) have identified very modest disease-associated methylation differences in MZ twins mapping to genes which either have known involvement in disease or are of biological relevance to the pathogenesis of disease (Dempster et al. 2011; Gervin et al. 2012; Rakyan et al. 2011a). These studies had small sample sizes of between 15-27 disease-discordant twin pairs. In contrast to previous studies that identified no sites that reached statistical significance, the current study identified 13 DMPs which surpassed a FDR of 5%. This may reflect the larger sample size available for testing, which enhanced the power to detect methylation differences.

Several of the DMPs identified in the current study occur at genetic loci previously associated with RA or within genes involved in pathways which are implicated in RA pathogenesis. Most notably, one of the most significantly associated DMPs (p=5.0E-6), which was hypomethylated in the RA twins in the cell corrected whole blood analysis, mapped to the zinc finger protein 74 (ZNF74) gene. The same probe was found to be differentially methylated in the same direction (p=5.6E-9) in the largest EWAS of RA (Liu et al. 2013). Interestingly, a study investigating gene expression in PBMCs found this gene was upregulated more than threefold in early RA (Olsen et al. 2004), indicating that the differential methylation could be altering gene expression in peripheral cells. The third most significantly DMP in whole blood unadjusted for cell composition (eleventh DMP when cell composition was adjusted for), which was hypermethylated in RA twins, maps to the protein kinase C, zeta (PRKCZ) gene and was previously found to be hypermethylated in RA FLS when compared to OA FLS (Nakano et al. 2013).

A probe mapping to the regulator of calcineurin 2 (RCAN2) gene was in the top ten DMPs in whole blood whether or not cell composition was adjusted for, and was hypermethylated in RA twins in both analyses. Interestingly this gene was
previously found to be hypermethylated in CD4+ T cells of JIA patients who had not been treated with methotrexate; however this effect was not found when methotrexate-treated individuals were included in analysis (Ellis et al. 2012).

The third most significant DMP in whole blood (corrected for cell composition) maps to the FOXK1 gene, which has previously been found to promote cell proliferation and regulate the transcription factor FoxO4 (Shi et al. 2012). Another member of the FoxO transcription factor family, FoxO1, is downregulated in RA FLS and PBMCs (Grabiec et al. 2014; Kuo and Lin 2007).

A probe mapping to the DICER1 gene was in the top 10 DMPs in whole blood with and without adjustment for cell composition. While this gene does not have an obvious biological link with RA, it encodes the enzyme Dicer which cleaves precursor RNA to produce miRNA indicating it is critical in the correct control of gene expression.

The 10th most significant DMP in whole blood uncorrected for cell composition mapped to the FC receptor-like 3 (FCRL3) gene. Polymorphisms in this gene have previously been associated with RA susceptibility risk in specific populations, including Iranian, Asian and Canadian, but not European populations (Eyre et al. 2006; Golmoghaddam et al. 2013; Kochi et al. 2005; Lee et al. 2010; Newman et al. 2006; Okada et al. 2014). The gene provides a good example of population stratification in RA susceptibility and shows the importance of considering different populations separately in genetic studies. My study suggests that the gene may be regulated by DNA methylation rather than genetic polymorphisms in European populations, which may explain why genetic studies have found this population stratification. Interestingly polymorphisms in this gene have also been associated with RA severity and disease activity rather than susceptibility in some populations (Bajpai et al. 2012; Han et al. 2012; Maehlen et al. 2011). These SNPs have also
been found to alter cell-surface expression of the FCRL3 protein in Treg cells in RA (Bajpai et al. 2012), which reinforces the importance of Treg cells in RA pathogenesis.

In the whole blood unadjusted analysis another differentially methylated gene was ST3GAL4 whose expression is controlled by TNF (Colomb et al. 2012; Colomb et al. 2014).

In the analyses of PBMC samples, very few DMPs were in regions previously associated with autoimmune disorders or inflammation. One DMP (in the analysis unadjusted for cell composition) mapped to the FBN1 gene which has been associated in genetic studies with JIA and systemic sclerosis (Tan et al. 2001; Wilson et al. 2013; Zhou et al. 2003). Another DMP in the same analysis mapped to the BCL2 gene which has previously been found to have increased expression in RA FLS when compared to OA FLS (Lee et al. 2013). In the cell corrected analysis, the 9th DMP mapped to the SIVA1 gene which negatively regulated NF-κB activity (Gudi et al. 2006).

The epigenetic signatures described in this study are limited by the fact that they could be either a cause of disease or a consequence of disease pathology. The elucidation of temporal origins of methylation differences could be investigated using longitudinal inception cohorts such as the ‘pre-clinical evaluation of novel targets in RA’ (PREVeNT RA) cohort currently being developed in Manchester. That study recruits first degree relatives of patients with RA, collecting clinical information and blood samples in healthy individuals then tracking their RA status longitudinally. While still in its infancy, the study has the potential to provide invaluable information about the development of RA and the temporal origins of methylation alterations. Identifying methylation that is causal for disease would allow elucidation of the functional impact of methylation changes and would be more informative of the
potential therapeutic or diagnostic impact in disease. Meanwhile, the current study provides further evidence of a role for DNA methylation in RA pathogenesis, which will act as a stepping stone in the investigation of disease pathogenesis in this complex disease.

4.3.3. **Effect of cryopreservation on methylation**

Due to the increasing number of studies in which cryopreservation of samples is undertaken prior to cell separation and DNA extraction, this study is of huge relevance and represents a finding which validates the reliability and suitability of the technique for epigenetic studies. While no significant changes were identified, small differences in methylation were observed indicating that in future, further investigation in a well powered study may identify more significant methylation changes. While it is undoubtable that cells are put under stress during cryopreservation, steps are taken to reduce the impact of the process on cellular viability, such as the addition of DMSO to the cryopreservation medium and the slow freezing/quick thawing technique, which both reduce the formation of crystals in the cells, making them less likely to burst while thawing. One observation which could influence future study design when cryopreserving cells was that a larger number of dead cells were observed in samples which had been cryopreserved compared to samples prepared while fresh. This may indicate that a larger sample is required to obtain enough live cells when cryopreserving the samples. The advantage of using FACS to sort the cells is that dead cells can be removed by gating. Another process for separating individual cell types, magnetic-activated cell sorting (MACS) relies on antibody selection and does not allow exclusion of dead cells, which should be considered during study design if cryopreserved cells are to be used.
4.4. **Future work**

4.4.1. **Replication of response to anti-TNF methylation differences**

In order to verify the results, it is important to perform replication studies in large independent cohorts of patients to ensure the DMPs were not false positives. Larger studies would also provide greater power to detect methylation differences. However, it is important to recognise that it has taken 5 years and recruitment across 57 NHS sites to recruit the patient samples tested in the current study. No other collections of this size are available world-wide in which to perform a replication study at this point in time. An alternative approach is to test the intermediate responders, which are larger in number. It is also important to investigate intermediate responders to biologic therapies, if the sensitivity and specificity of the markers identified are to be assessed for their ability to predict response in all patients starting anti-TNF therapy. While replication of these findings using genome-wide methylation analysis such as the HumanMethylation450 BeadChip described in this thesis, or ideally whole genome bisulphite sequencing, would be most informative and more likely to provide information about novel methylation sites, these techniques are prohibitively expensive and so would reduce the feasible sample size for replication studies. Replication studies which target the differentially methylated sites identified in the current studies could be performed using much cheaper gene-targeted bisulphite sequencing, such as pyrosequencing. Power calculations should be performed to estimate the number of good and poor responders who should be included in the replication cohort to provide sufficient power to detect methylation differences to a degree that would allow translation of the methylation signature into a clinical test. The lower cost of these techniques would allow a greater number of samples to be included in the replication cohort, and hence a greater power in the study.
The response studies in this thesis focussed on etanercept and adalimumab, the most commonly prescribed anti-TNF biologic therapies. In future studies it would be interesting to investigate response to other anti-TNFs such as certolizumab and golimumab. It is also important to investigate non-biologic therapies in RA which have poor response rates, for example methotrexate, which is the most commonly prescribed DMARD.

Treatment response is a very complex phenotype, and the methylation differences identified in this thesis are likely to contribute a small amount to the variability of that phenotype. To date, genetic studies have also identified a genetic component in treatment response, with limited power, alongside environmental components such as smoking. Due to the complexity of the trait, it is unlikely that a clinical test of response will rely on just one of these components. In order to provide a prediction of treatment response in clinic, I would suggest the development of a composite algorithm, combining clinical and lifestyle factors known to influence treatment response such as baseline DAS28, gender and smoking status with biomarkers predictive of response such as the methylation marks identified in this thesis, and SNP markers associated with treatment response. By estimating the relative contribution of each of these factors to treatment response, each component in the algorithm could be given a weighted score to allow prediction of treatment response to a given therapy in clinic. As with other composite clinical measurements, such as DAS28, this is likely to have some measurement bias and so would have to be developed with caution and preferentially would include as few subjective measures as possible, to prevent the interpretation bias seen in DAS28.

The ‘Maximising Therapeutic Utility for Rheumatoid Arthritis’ (MATURA) programme, based in Manchester, is working towards stratified medicine by aiming to identify epigenetic, transcriptomic and proteomic biomarkers of response to RA
therapies. By collecting pre-treatment samples, this is an ideal cohort for use in the identification of epigenetic predictors of treatment response.
4.4.2. Investigation into alternative predictors of anti-TNF treatment response

Studies investigating alternative predictors of anti-TNF treatment response such as other classes of epigenetic marks, transcriptomics and proteomics, could potentially allow the identification of novel markers of treatment response. They may also have the potential to elucidate the functional mechanism by which existing markers of treatment response affect the complex phenotype. This approach of measuring multiple omics is even more powerful when the experiments can be performed in the same individuals. A first step in combining different omics approaches would be to investigate the interaction between methylation differences and genetic background. For example methylation quantitative trait loci (meQTLs) could be identified by combining SNP and DMP data, to determine whether specific methylation sites correlate directly with genetic variants. Following on from the identification of methylation differences in treatment response, another important step is to investigate gene expression. Measuring gene expression (preferably in the same individuals), when combined with methylation data generated in this experiment, would allow an exploration of whether expression of a gene correlates with methylation, however this does not necessarily indicate causality, ie methylation does not necessarily directly control gene expression.

Once methylation and expression are known to correlate, the direction of the causal relationship could be determined using in vitro experiments. Recent development of epigenome-altering techniques (such as the TALE-TET1 fusion proteins) could be employed to directly demethylate a gene of interest in cell culture and measure the effect on gene expression (Maeder et al. 2013). There have been conflicting results about the effect of methylation on gene expression, and an experiment such as this would be the most definitive way to measure the downstream functional relevance of a methylation change. In the future, such techniques could even be employed for
therapeutic benefit by altering the methylation, and hence the expression, of disease-causing or treatment response-inhibiting genes.

Investigation into histone modifications may also provide novel markers of treatment response; however the practicalities of performing such experiments make the translation of such biomarkers to clinic unfeasible. While their potential as a clinical biomarker is currently limited, chromatin immunoprecipitation (ChIP) assays are useful for the functional characterisation of SNP and DMP markers. These assays can identify if markers disrupt binding of DNA to a histone which may affect activation and hence expression of the associated gene. Identification of the histone marks which overlap a particular genetic or DNA methylation mark can also be informative of the role of the DNA sequence in question, for example enhancers of genes typically overlap the monomethylation of histone H3 at lysine 4 (H3K4me1) and acetylation of histone H3 at lysine 27 (H3K27ac) histone modifications (Shlyueva, Stampfel, and Stark 2014). Histone modifications can also be cell type specific, for example the trimethylation of histone H3 at lysine 4 (H3K4me3) mark which is an indicator of active gene regulation (Trynka et al. 2013). By investigating if the genetic and epigenetic marks of treatment response overlap with such histone modifications, the cell type which is most prolific in driving treatment response could be identified. This would inform future studies about which cell type is most important to focus on in the investigation of treatment response.

Due to the difficulties in dealing with the composite nature of the DAS28 measurement of treatment response, which includes both objective and subjective measures, it is also important to investigate methylation differences using the subcomponents of DAS28 such as CRP levels and swollen joint counts (Plant, Wilson, and Barton 2014). These measurements are more objective, so it would be interesting to see if the methylation differences are greater when using these instead of DAS28 as a measure of treatment response.
4.4.3. **Investigation into the mechanism by which LRPAP1 methylation could be affecting anti-TNF response**

The studies described in this thesis will act as a launchpad for larger scale studies investigating methylation as a biomarker of response, but it is also important to investigate the functional role of the methylation changes identified in this study. A possible next step is to perform a hypothesis-led study to test the proposed mechanism by which differential methylation of the LRPAP1 gene could be affecting response to etanercept (described in section 4.3.1.1). First of all it would be important to measure the change in expression caused by the methylation differences, and to see if this has downstream effects by measuring protein levels of LRP1 and LRPAP1 in responders and non-responders to etanercept. I hypothesised that increased levels of LRP1 were inhibiting response to etanercept in non-responders. This could be tested by culturing RA FLS and Treg cells and treating with etanercept, with and without co-treatment with an LRP1 antagonist. It would also be very interesting to test whether genetic variants are mediating methylation differences by correlating underlying SNPs in the regions of DMPs, in order to identify meQTLs.
4.4.4. Establishing the temporal origins of rheumatoid arthritis methylation changes

In order to establish the temporal origins of methylation changes in RA it is necessary to look at longitudinal studies such as the PReVENT RA study described in section 4.3.2. Such studies allow investigation into pre-diagnosis RA DNA methylation which would be informative of whether the methylation changes identified in this study (and other case-control studies of established RA) are causative for RA or a consequence of disease pathology. A limitation of many studies to date is that they have collected whole blood DNA, while investigation into individual cell types may be more appropriate and allow identification of methylation differences which are specific to certain cell populations and which may be missed in analysis of whole blood. Sample collections can take an extensive amount of time to grow, and currently there is not one defined cell type in which methylation would be most useful to investigate, thus the most appropriate measure to take at the moment for future epigenetic studies is to collect PBMCs and cryopreserve them to allow isolation of particular cell types of choice in the future. The cryopreservation study in the current thesis has shown that this is an appropriate technique for epigenetic studies and should not influence the DNA methylation extensively.
4.4.5. Characterising the functional effects of methylation changes in rheumatoid arthritis

In order to elucidate the functional consequences of methylation changes in RA identified from the current twin study, it is important to measure the downstream effects such as altered expression in these genes. It would also be interesting to see if the methylation changes are interacting with other epigenetic phenomena such as histone modifications which may alter chromatin conformation, affecting expression of genes. This may be particularly interesting in the DMPs identified in intergenic regions, as their role is currently unclear.

It would also be interesting to investigate the relationship between disease-associated genetic marks and the methylation differences identified in this study. Analysing the discordant twin methylation data described in this thesis enriching for the genetic RA susceptibility loci, subtle methylation differences associated with these loci could be investigated. By combining the methylation data with genetic data, meQTLs could also be identified which could elucidate the relationship between the two modifications and disease. This could allow the identification of a mechanism by which methylation status controls the impact of genetic signatures in disease by mediating the expression of the associated genes. Due to the increased occurrence of RA in the female population, it would also be interesting to perform the analysis on the subset of probes which map to the X-chromosome, just in female participants, to see if X-linked methylation differs between disease discordant twins.
4.5. **Conclusion**

A methylation signature has been identified that indicates response to etanercept. No significant methylation differences were identified in relation to response to adalimumab, however this may be due to a conservative sample size. Further genome-wide investigations of methylation as a biomarker of anti-TNF therapy response are warranted, and a targeted gene validation study of the markers identified in this study is necessary. Subtle differences in methylation were identified in RA discordant monozygotic twin pairs; however these were not statistically significant once cell composition was adjusted for. No differences in DNA methylation were identified between cells which had been prepared from cryopreserved or fresh samples, indicating cryopreservation of cells is an appropriate technique for use in epigenetic studies.


Heijmans, B. "The prenatal epigenome and studies into long-term health", in Epigenomics of Common Disease Conference.


Hinks, A., Barton, A., John, S., Bruce, I., Hawkins, C., Griffiths, C. E. et al. (2005). Association between the PTPN22 gene and rheumatoid arthritis and juvenile idiopathic arthritis in a UK population: further support that PTPN22 is an autoimmunity gene. Arthritis Rheum, 52, (6), 1694-1699.


Potter, C., Cordell, H. J., Barton, A., Daly, A. K., Hyrich, K. L., Mann, D. A. *et al.* (2010). Association between anti-tumour necrosis factor treatment response and


Strangfeld, A., Eveslage, M., Kekow, J., Grabetaler, A., Kaufmann, J., Listing, J. *et al.* (2009). Effectiveness of treatment with rituximab depends on autoantibody status results from 2 years of experience in the German biologics register RABBIT. *Arthritis Rheum*, **60**.


Svendsen, A. J., Gervin, K., Lyle, R., Nielsen, C., Houen, G., Junker, P., Kyvik, K., & Tan, Q. 2014, "Changes in DNA methylation related to smoking are associated with RA. An epigenome-wide association study in monozygotic RA discordant twin pairs.", in *EULAR*.


Zymo Research (2012a), *EZ DNA Methylation Kit Instruction Manual Version 1.2.2.*