Post-GWAS Bioinformatics and Functional Analysis of Disease Susceptibility Loci

A thesis submitted to The University of Manchester for the degree of Doctor of Philosophy (PhD) in the Faculty of Biology, Medicine and Health

2016

Paul Martin

Table of Contents

Table of Figures

Table of Tables

List of Abbreviations

3C	chromosome conformation capture
4C	chromosome conformation capture-on-chip
5C	chromosome conformation capture carbon copy
ACPA	anti-cyclic citrullinated peptide antigens
AID	autoimmune disease
APC	antigen presentiing cell
API	application programming interface
AS	ankylosing spondylitis
ATD	autoimmune thyroid disease
bp	base pair
CADD	combined annotation-dependent depletion
CCP	cyclic citrullinated peptides
CEEHRC	Canadian Epigenetics, Environment and Health Research Consortium
CEL	coeliac disease
CGI	Common Gateway Interface
ChIP-exo	chromatin immunoprecipitation combined with lambda exonuclease digestion followed by high-throughput sequencing
ChIP-Seq	chromatin immunoprecipitation followed by next-generation sequencing
CREST	Core Research for Evolutional Science and Technology
CRISPR	clustered regularly interspaced short palindromic repeats
CRO	crohn’s disease
CRP	C-reactive protein
CTLA-4	cytotoxic T-lymphocyte-associated protein 4
CyTOF	cytometry by time of flight (CyTOF)
DBI	database interface
DEEP	Deutsches Epigenom Programm
DMARD	disease-modifying anti-rheumatic drug
DNA	deoxyribonucleic acid
dsQTL	Dnase I sensitivity quantitative trait loci
ENCODE	ENCyclopaedia Of DNA Elements
eQTL	expression quantitative trait loci
ESC	embryonic stem cell
EULAR	European League Against Rheumatism
FLC	foetal liver cell
GoShifter	Genomic Annotation Shifter
GRAIL	Gene Relationships Across Implicated Loci

GTEx	Genotype-Tissue Expression
GWAS	Genome-wide association study
GxE	Gene–environment interaction
HLA	human leukocyte antigen
HS	hypersensitivity site
IBD	inflammatory bowel disease
IFN	interferon
IHEC	International Human Epigenome Consortium
IL-1	interleukin 1
IL-17	interleukin 17
IL-2	interleukin 2
IL-6R	interleukin 6 receptor
INSIGHT	Interspersed Genomically coHerent elemenTs
JAK	Janus kinase
JDBC	java database connectivity
JIA	juvenile idiopathic arthritis
JSP	JavaServer pages
kb	kilobase
LCL	lymphoblastoid cell line
LCR	locus control region
LD	linkage disequilibrium
LDL-C	low-density lipoprotein cholesterol
lncRNA	long non-coding RNA
LPS	lipopolysaccharide
Mb	megabase
MRC	Medical Research Council
MS	multiple sclerosis
NEGEG	North of England Genetic Epidemiology Group
NGS	next generation sequencing
NHS	National Health Service
nm	nanometres
OD	other diseases
OR	odds ratio
PBC	primary biliary cirrhosis
PICS	Probabilistic Identification of Causal SNP
PPI	protein-protein interaction
Ps	psoriasis

PsA	psoriatic arthritis
PWM	position weight matrix
RA	rheumatoid arthritis
RNA	ribonucleic acid
RR	relative risk
SBS	sequence by synthesis
SINE	short interspersed element
SLE	systemic lupus erythematosus
SNP	single nucleotide polymorphism
SOLiD	small oligonucleotide ligation and detection
STAT	Signal Transducer and Activator of Transcription
TID	type 1 diabetes
T2C	Targeted Chromatin Capture
TAD	topologically associated domain
TCR	T-cell receptor
TF	transcription factor
T_H17	$CD4^+$ type 17 T helper cell
TNF	tumour necrosis factor
TLR	Toll-like receptor
Treg	regulatory T-cell
TSS	transcription start site
UC	ulcerative colitis
UCSC	Univeristy of California, Santa Cruz
UK	United Kingdom
UTR	untranslated region
VDR	vitamin D receptor
VDRE	vitamin D response element
VEP	Variant Effect Predictor
$VitD_3$	1,25-dihydroxyvitamin D3
WT ISSF	Wellcome Trust Institutional Strategic Support Fund
WTCCC	Wellcome Trust Case Control Consortium
XML	extensible markup language

Abstract

Genome-wide association studies (GWAS) have been tremendously successful in identifying genetic variants associated with complex diseases, such as rheumatoid arthritis (RA). However, the majority of these associations lie outside traditional protein coding regions and do not necessarily represent the causal effect. Therefore, the challenges post-GWAS are to identify causal variants, link them to target genes and explore the functional mechanisms involved in disease. The aim of the work presented here is to use high level bioinformatics to help address these challenges.

There is now an increasing amount of experimental data generated by several large consortia with the aim of characterising the non-coding regions of the human genome, which has the ability to refine and prioritise genetic associations. However, whilst being publicly available, manually mining and utilising it to full effect can be prohibitive. I developed an automated tool, ASSIMILATOR, which quickly and effectively facilitated the mining and rapid interpretation of this data, inferring the likely functional consequence of variants and informing further investigation. This was used in a large extended GWAS in RA which assessed the functional impact of associated variants at the 22q12 locus, showing evidence that they could affect gene regulation.

Environmental factors, such as vitamin D, can also affect gene regulation, increasing the risk of disease but are generally not incorporated into most GWAS. Vitamin D deficiency is common in RA and can regulate genes through vitamin D response elements (VDREs). I interrogated a large, publicly available VDRE ChIP-Seq dataset using a permutation testing approach to test for VDRE enrichment in RA loci. This study was the first comprehensive analysis of VDREs and RA associated variants and showed that they are enriched for VDREs, suggesting an involvement of vitamin D in RA.

Indeed, evidence suggests that disease associated variants effect gene regulation through enhancer elements. These can act over large distances through physical interactions. A newly developed technique, Capture Hi-C, was used to identify regions of the genome which physically interact with associated variants for four autoimmune diseases. This study showed the complex physical interactions between genetic elements, which could be mediated by regions associated with disease. This work is pivotal in fully characterising genetic associations and determining their effect on disease. Further work has re-defined the 6q23 locus, a region associated with multiple diseases, resulting in a major re-evaluation of the likely causal gene in RA from *TNFAIP3* to *IL20RA*, a druggable target, illustrating the huge potential of this research. Furthermore, it has been used to study the genetic associations unique to multiple sclerosis in the same region, showing chromatin interactions which support previously implicated genes and identify novel candidates. This could help improve our understanding and treatment of the disease.

Bioinformatics is fundamental to fully exploit new and existing datasets and has made many positive impacts on our understanding of complex disease. This empowers researchers to fully explore disease aetiology and to further the discovery of new therapies.

Declaration

The University of Manchester

PhD by published work Candidate Declaration

Candidate Name: Paul Martin

Faculty: Faculty of Biology, Medicine and Health

Thesis Title: Post-GWAS Bioinformatics and Functional Analysis of Disease Susceptibility Loci

1. I am first or joint first author of publications 1, 3, 4 and 6 and prominently involved in publications 2 and 5, making contributions to the design, execution and analysis of experiments and drafting and submitting manuscripts.

Publication 1:
P. Martin, A. Barton & S. Eyre. ASSIMILATOR: a new tool to inform selection of associated genetic variants for functional studies. *Bioinformatics.* **27**(1), 144-146 (2011).

This paper describes a bespoke bioinformatics tool to efficiently mine a large and expanding publicly available dataset. I recognised that whilst this data was essential to integrate into the genetic work of the department, it was difficult to access the complete dataset efficiently. In consultation with other members of the group, I developed a tool (ASSIMILATOR) for researchers within the department, although it soon became apparent that it would also have an application to the wider genetics community. I planned and developed the tool, taking on board suggestions from the team, and extended it for online use to make it more accessible to researchers both internal and external to the university. I was fully responsible for the writing and submission of the manuscript.

Publication 2:
G. Orozco, S. Viatte, J. Bowes, **P. Martin**, A. G. Wilson, A. W. Morgan, S. Steer, P. Wordsworth, L. J. Hocking, A. Barton, J. Worthington & S. Eyre. Novel rheumatoid arthritis susceptibility locus at 22q12 identified in an extended UK genome-wide association study. *Arthritis Rheumatol.* **66**, 24–30 (2014).

This work presents the results of an extended GWAS in RA which was followed by additional bioinformatics analyses including the utilisation of the ASSIMILATOR programme. I was involved in the analysis and interpretation of the data, primarily the bioinformatics analysis and also contributed to the preparation of the manuscript.

Publication 3:

A. Yarwood*, **P. Martin***, J. Bowes, M. Lunt, J. Worthington, A. Barton & S. Eyre. Enrichment of vitamin D response elements in RA-associated loci supports a role for vitamin D in the pathogenesis of RA. *Genes Immun.* **14**(5), 325-329 (2013). *joint first author

The role of vitamin D in RA has previously been established. In this paper I am joint first author and utilised a publicly available ChIP-Seq dataset on VDREs and developed a custom method to show that RA associated loci are enriched for VDREs further supporting the role of vitamin D in the development of RA. I was involved in initial discussions regarding the hypothesis of the study and the aim of the study. I then downloaded the publicly available data in the most appropriate format and developed a script to simultaneously process this data and test for enrichment of VDREs in RA associated loci and 100,000 random control datasets. The results were then compiled and the manuscript prepared with the joint lead author who contributed the RA genetic association data.

Publication 4:

P. Martin*, A. McGovern*, G. Orozco*, K. Duffus, A. Yarwood, S. Schoenfelder, N. Cooper, A. Barton, C. Wallace, P. Fraser, J. Worthington & S. Eyre. Capture Hi-C reveals novel candidate genes and complex long-range interactions with related autoimmune risk loci. *Nat. Comms.* **6**:10069 doi: 10.1038/ncomms10069 (2015). *joint first author

The aim of this study was to use Capture Hi-C to characterise the physical interactions affecting gene regulation in four autoimmune diseases. I contributed to the initial discussions regarding the hypothesis and aims of the study. This led to the overall concept and design of the work presented in the paper. Subsequently I designed the experiments, developed a pipeline to quickly and easily design RNA capture baits and liaised with Agilent during their manufacture. I had regular discussions with co-authors who performed the experimental work and organised the sequencing of the libraries. I contacted core facilities to obtain the raw data and designed and performed several subsequent analysis steps to fully analyse and interpret the data. I prepared and co-ordinated the revision of the manuscript, including all figures and submitted the paper.

Publication 5:

A. McGovern, S. Schoenfelder, **P. Martin**, J. Massey, K. Duffus, D. Plant, A. G. Pratt, A. E. Anderson, J. D. Isaacs, J. Diboll, N. Thalayasingam, C. Ospelt, P. Fraser, A. Barton, J. Worthington, S. Eyre & G. Orozco. Capture Hi-C identifies a novel causal gene, *IL20RA*, in the pan-autoimmune genetic susceptibility region 6q23. *Genome Biol.* **17**, 212 (2016)

This paper builds on the previous Capture Hi-C study (Publication 4) by further examining the 6q23 locus associated with several autoimmune diseases including RA. Following validation of the interactions identified in the Capture Hi-C study, it showed that disease risk variants effect the expression of *IL20RA* and provided evidence for the efficacy of an anti-IL-20 therapy effective in the treatment of RA. I was involved in identifying the interactions to follow-up, the bioinformatics analysis and preparing the manuscript.

Publication 6:

P. Martin*, A. Mcgovern*, J. Massey, S. Schoenfelder, K. Duffus, A. Yarwood, A. Barton, J Worthington, P. Fraser, S. Eyre & G. Orozco. Identifying Causal Genes at the Multiple Sclerosis Associated Region 6q23 Using Capture Hi-C. *PLoS ONE*. **11**, e0166923 (2016) *joint first author

This paper exploits the finding that many loci are shared between autoimmune diseases to identify the target genes for variations uniquely associated with multiple sclerosis (MS) in the 6q23 region by investigating chromatin interaction data from our previous Capture Hi-C study (Publication 4). In addition, associated variants were refined using publicly available data on regulatory elements. I conducted the analysis, interpreted the results and prepared the manuscript. I co-ordinated revision of the manuscript and submitted the paper.

2. All of the work presented here was completed whilst I have been a member of staff at The University of Manchester.
3. No portion of the work completed by me, outlined in this declaration and referred to in this 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.

I confirm that this is a true statement and that, subject to any comments above, the submission is my own original work

Signed:……………………………… **Date:……………………………………**

Copyright Statement

Acknowledgments

I would like to thank all those who contributed to the publications presented in this thesis. Thanks also to Arthritis Research UK for providing me with the funding and opportunity to conduct this research, my supervisor, Dr Steve Eyre and Dr Gisela Orozco, for their help and input throughout. Finally I would like to thank my wife Nicola for her help and support.

Statement

i. Particulars of the candidate's degrees, other qualifications and research experience:

Qualifications:

2011 MPhil Medicine, The University of Manchester

2004 BSc (Hons) Molecular Biology with Industrial Experience 2ii, The University of Manchester

Positions held:

2011-present Research Associate, Arthritis Research UK Centre of Excellence for Genetics and Genomics, University of Manchester

2007-2011 Research Assistant, Arthritis Research UK Epidemiology Unit, University of Manchester

Experience:

In the years since moving to the University of Manchester from the Sanger Institute, Cambridge, my work has contributed to the understanding of genetic association to complex disease; prioritisation of likely causal variants based on available functional data and identification of causal genes through nuclear dynamics (chromosome conformation capture) studies. My research has contributed to the understanding of the genetic basis of complex diseases through the integration and utilisation of large, diverse datasets from both multiple public international initiatives and data generated from within the Manchester group. I have a total of 29 peer reviewed publications, including ten in top rheumatology journals, one Nature and four Nature Genetics publications.

Conferences:

Poster: “Chromatin Interactions Reveal Novel Gene Targets for Drug Repositioning in Rheumatic Diseases”, American College of Rheumatology Annual Meeting, Washington D.C., November 2016

Platform (Oral Presentation): “Capture Hi-C identifies compelling candidate causal genes and enhancers for multiple sclerosis in the 6q23 region”, 66th Annual Meeting of The American Society of Human Genetics, Vancouver, October 2016

Poster: “Capture Hi-C identifies compelling candidate causal genes and enhancers for multiple sclerosis in the 6q23 region”, European Journal of Human Genetics Conference, Barcelona, May 2016

Platform (Oral Presentation): “Chromosome interaction analysis of risk loci in related autoimmune diseases reveals complex, long-range promoter interactions implicating novel candidate genes”, 65th Annual Meeting of The American Society of Human Genetics, Baltimore, October 2015

Attended: Genome Science: Biology, Technology & Bioinformatics 2014, Oxford, September 2014.

Poster: “Differences and Overlap of Immunological Pathways Implicated in the Aetiology of Anti-Citrullinated Peptide Antibody Positive and Negative Rheumatoid Arthritis”, American College of Rheumatology Annual Meeting, San Diego, October 2013.

Poster Tour: “Comparison of pathways implicated in anti-citrillunated peptide antibody positive and negative rheumatoid arthritis patients”, EULAR Congress, Madrid, June 2013.

Poster: “Comparison of pathways implicated in anti-citrillunated peptide antibody positive and negative rheumatoid arthritis patients”, 62nd Annual Meeting of The American Society of Human Genetics, San Francisco, November 2012.

Presentation: “ASSIMILATOR: a new tool to inform selection of associated genetic variants for functional studies”, North of England Genetic Epidemiology Group (NEGEG) meeting, Leeds, November 2010.

Awards:

Poster award - European Journal of Human Genetics Conference, Barcelona, May 2016

Key Research Skills:

Extensive knowledge of the Perl programming language including CGI, Perl/Tk and the Perl DBI and also knowledge of higher level programming languages such as Java, including the JDBC API and JSP and C.

Extensive knowledge of Oracle and MySQL database use and development.

Extensive experience using statistical packages such as R.

Extensive knowledge of Microsoft Office and the Linux based Open Office as well as Windows and Linux based operating systems.

Independent problem solving and management of multiple projects in a timely manner.

Ability to develop research concepts and work with other group members to create novel solutions.

Experience in the analysis of next generation sequencing data, including Illumina sequencing by synthesis (SBS), Life Technologies SOLiD and Roche 454, for multiple applications such as RNA-Seq, ChIP-Seq & Hi-C.

Ability to utilise existing skills to design experiments which maximise research potential whilst minimising costs.

Preparation of manuscripts and scientific data for publication in a clear and concise format.

Grants:

"The nanoscale organisation of naïve and regulatory T cell surfaces in juvenile idiopathic arthritis patients", MRC Discovery Award: 'Capacity Building in single cell inflammation discovery: developing the next generation of scientists', co-applicant, 2016 (£100k Awarded)

"Monocyte heterogeneity in health and inflammatory disease", MRC Discovery Award: 'Capacity Building in single cell inflammation discovery: developing the next generation of scientists', co-applicant, 2016 (£73,326 Awarded)

"Translating findings from genome wide association studies into novel therapeutic targets for rheumatoid arthritis", Arthritis Research UK Project Grant, co-applicant, 2016 (£195,835 Awarded)

Wellcome Trust Institutional Strategic Support Fund (WT ISSF), Strategic Awards in Single Cell Research, co-applicant, 2015 (£12k Awarded)

ii. A complete and numbered list of the publications submitted (grouped according to subject and type)

Publication 1:

P. Martin, A. Barton & S. Eyre. ASSIMILATOR: a new tool to inform selection of associated genetic variants for functional studies. *Bioinformatics.* **27**(1), 144-146 (2011).

Publication 2:

G. Orozco, S. Viatte, J. Bowes, **P. Martin**, A. G. Wilson, A. W. Morgan, S. Steer, P. Wordsworth, L. J. Hocking, A. Barton, J. Worthington & S. Eyre. Novel rheumatoid arthritis susceptibility locus at 22q12 identified in an extended UK genome-wide association study. *Arthritis Rheumatol.* **66**, 24–30 (2014).

Publication 3:

A. Yarwood*, **P. Martin***, J. Bowes, M. Lunt, J. Worthington, A. Barton & S. Eyre. Enrichment of vitamin D response elements in RA-associated loci supports a role for vitamin D in the pathogenesis of RA. *Genes Immun.* **14**(5), 325-329 (2013). *joint first author

Publication 4:

P. Martin*, A. McGovern*, G. Orozco*, K. Duffus, A. Yarwood, S. Schoenfelder, N. Cooper, A. Barton, C. Wallace, P. Fraser, J. Worthington & S. Eyre. Capture Hi-C reveals novel candidate genes and complex long-range interactions with related autoimmune risk loci. *Nat. Comms.* **6**:10069 doi: 10.1038/ncomms10069 (2015). *joint first author

Publication 5:

A. McGovern, S. Schoenfelder, **P. Martin**, J. Massey, K. Duffus, D. Plant, A. G. Pratt, A. E. Anderson, J. D. Isaacs, J. Diboll, N. Thalayasingam, C. Ospelt, P. Fraser, A. Barton, J. Worthington, S. Eyre & G. Orozco. Capture Hi-C identifies a novel causal gene, *IL20RA*, in the pan-autoimmune genetic susceptibility region 6q23. *Genome Biol.* **17,** 212 (2016).

Publication 6:

P.Martin*, A. Mcgovern*, J. Massey, S. Schoenfelder, K. Duffus, A. Yarwood, A. Barton, J Worthington, P. Fraser, S. Eyre & G. Orozco. Identifying Causal Genes at the Multiple Sclerosis Associated Region 6q23 Using Capture Hi-C. *PLoS ONE*. **11**, e0166923 (2016). *joint first author

iii. An overall summary of the aims and achievement of the work, for which the publications submitted give evidence.

1 Introduction

1.1 Complex Disease Genetics

Many common diseases, such as type 1 diabetes (T1D), inflammatory bowel disease (IBD) and rheumatoid arthritis (RA), are caused by a complex combination of genetic, environmental and lifestyle factors. Historically, complex disease genetics were investigated using linkage or candidate gene studies. While these approaches were successful in identifying genetic changes and genes causing monogenic disorders[1,2], they have had limited application to complex disorders as, with a few notable exceptions with large effect sizes[3–5], they were underpowered to detect much of the genetic susceptibility to disease.

More recently, since 2007 starting with the Wellcome Trust Case Control Consortium (WTCCC)[6], the modern complex disease genetics era have utilised genome-wide association studies (GWAS). GWAS have been tremendously successful in identifying genetic variants associated with common complex diseases in a hypothesis-free way[6–11] and were made possible by technological developments in array based genotyping methods, pioneered by Affymetrix and Illumina. They compare the allele frequency of thousands of markers across the whole genome, usually single nucleotide polymorphisms (SNPs), between cases (individuals with disease) and controls (healthy individuals) to determine if one allele occurs in cases more or less often than expected; the statistical significance of this difference is then determined.

Due to the inherited nature of the genome, the vast majority of associations identified in GWAS only provide the initial signposts for the identification of the genetic variants underpinning susceptibility to disease and do not represent the causal effect but one in linkage disequilibrium (LD). LD describes how alleles can segregate based on recombination during meiosis; alleles of SNPs in high LD are found together more often than SNPs in low LD. As such, the causal variation could be any SNP that lies in high LD with the GWAS associated SNP and can often implicate large genomic regions representing many potential causal SNPs. It can therefore be difficult to pinpoint this association to a specific region or gene. Additionally, for many complex diseases, the vast majority of the genetic associations identified are found outside traditional protein coding genes and recent studies have shown that they are enriched in enhancer elements suggesting that they are involved in gene regulation[12–14].

In an attempt to refine or ‘fine map’ the genetic associations identified, coupled with the insight that many common complex disease associations share significant overlap with each other, the Immunochip consortium sought to developed a genotyping array which could achieve this in a cost effective way to allow much larger sample numbers[15]. This approach not only fine mapped many associations but was also successful in identifying new variants associated with disease[16–27]. Although the Immunochip array, and subsequent large imputed meta-analyses[28] were successful, it was still not possible to fully resolve the causal variations underpinning disease susceptibility. It is therefore clear that identifying the causal SNP and, more importantly, the underlying disease mechanisms using GWAS/genetic evidence alone is typically not possible and considerable challenges remain if we are to fully translate GWAS findings into an understanding of disease aetiology.

Therefore, the challenges post-GWAS are firstly to determine which of these variants is most likely causal, secondly, which gene they regulate and finally, how the disease associated allele affects the functional mechanisms involved in disease. The basis of my work contributing to this thesis has been to use high level bioinformatics to help address these challenges.

Specifically, in the first publication[29], I developed a one-stop solution that quickly and effectively allowed genetics researchers to mine and rapidly interpret the data generated by the ENCyclopaedia Of DNA Elements (ENCODE) project[30] with ease. This has enabled researchers to easily identify and prioritise potential causal candidate variants for further investigation. This required multiple bioinformatics skills, including expertise in programming and databases and coupled with my background in molecular biology allowed me to develop a tool that researchers could use to fully utilise this resource. This was accomplished by developing a web-based interface to allow researchers both internally and externally to access and use the tool efficiently. This has been used in multiple publications to assess and prioritise genetic variants associated with disease. For example, the second paper[31] describes a large extended GWAS in RA which used this tool to assess the functional impact of variants associated at the 22q12 locus. Evidence was discovered suggesting an associated variant, rs1043099, and correlated variants map to sites of transcription factor binding and open chromatin. Coupled with histone modification evidence, this suggests that these associated variants could affect gene regulation.

In the third paper[32], I utilised a large publicly available dataset comprising of vitamin D response element (VDRE) ChIP-Seq data[33] and incorporated it with our existing genetic evidence. This custom analysis involved processing and filtering the data into a useable format and utilised a permutation testing approach to test for VDRE enrichment in RA loci and matched random controls. This study showed that variants associated with RA are enriched for VDREs, providing a link between vitamin D, a non-genetic factor, and RA. This study was the first comprehensive analysis of VDREs and RA associated variants and provides evidence for in involvement of vitamin D in RA and has the potential to inform research into vitamin D therapy in RA.

The fourth paper[34] represents a large study that was developed and carried out in Manchester to infer causal genes from genetic associations for four autoimmune diseases. This unique study has shown the complex physical interactions between genetic elements which exist in the nucleus and are mediated by regions associated with disease. This work involved careful consideration with regards to experimental design and subsequent analysis as it was one of the first to employ the Capture Hi-C technology and utilised a unique study design. The sequence data generated by this study was roughly equivalent to the amount required for three human genomes and required the development of custom pipelines and analyses to identify the multiple complex effects observed. Furthermore, thousands of interactions were identified and the results had to be stringently filtered to provide robust, validated interactions which had strong biological effect. This study is, and will continue to be, pivotal in subsequent functional experiments to fully characterise genetic associations and determine their effect on disease.

Although only recently published, the technique has generated considerable interest from researchers in a number of areas and is already being applied to post-GWAS investigation of a number of different diseases. This has already led to further work by our group, presented in the final two papers[35,36], the first of which has re-defined a genomic region associated with multiple diseases. An in depth analysis of the interactions of the 6q23 locus has resulted in a major re-evaluation of the likely causal gene from *TNFAIP3* to *IL20RA*, a drugable target. The final paper investigated the interactions involving variants only associated with multiple sclerosis (MS) in the 6q23 region. This showed that MS associated variants are involved in two clusters of interactions: one containing neurologically related genes and the other immunologically related genes, showing that individual variants could regulate multiple genes and that multiple independent variants could co-regulate groups of functionally similar genes. These two papers illustrate the huge potential impact of this and similar subsequent research.

1.2 Rheumatoid Arthritis

Rheumatoid arthritis (RA) is a common complex autoimmune disease characterised by chronic inflammation of the synovial joints leading to irreversible joint damage, disability and increased mortality. It is the most common form of arthritis in the UK, affecting approximately 1% of the population worldwide and costs the NHS around £3.5 billion per annum. However, the cause of RA is still unknown and current treatment options are not always effective. As such, it is important to understand which factors contribute to an individual's risk to RA to allow clinicians to effectively manage disease. The largest predisposing factor for developing RA is the genetic background of an individual, with λ_s estimates ranging from 5-10[37], and genetic association studies have been successful in identifying over 100 genetic regions containing variants associated with RA[6,10,11,19,28,38–40].

1.2.1 RA Genetic Associations

The largest and first genetic effect identified comes from the human leukocyte antigen (HLA) region and the class II *HLA-DRB1* gene, specifically copies of the *HLA-DRB1* gene containing the shared epitope, a five amino acid motif which confers susceptibility to RA (amino acid positions 70-74)[41]. More recently, a new model for the association of the HLA region in RA has been proposed which has identified that five amino acids in three HLA proteins (HLA-DRβ1, HLA-B & HLA-DPβ1) can explain most of the HLA risk in RA patients who have antibodies against cyclic citrullinated peptides (anti-CCP-positive RA)[42]. The main finding of this study was that HLA-DRβ1 risk could be defined by three amino acids at positions 11, 71 and 74 which, whilst offering new insights into RA HLA association does not radically alter the existing shared epitope hypothesis. A recent study in anti-CCP-positive ($ACPA^+$) RA, has also shown that amino acids at these positions are associated with severity, mortality and treatment response in RA patients[43]. Additionally, a study in anti-CCP-negative ($ACPA^-$) RA showed that while HLA-DRβ1 is associated with this subtype, albeit with a lower effect size, different HLA-DRβ1 alleles also have a different direction of effect (i.e. risk vs protective) (Figure 1)[44].

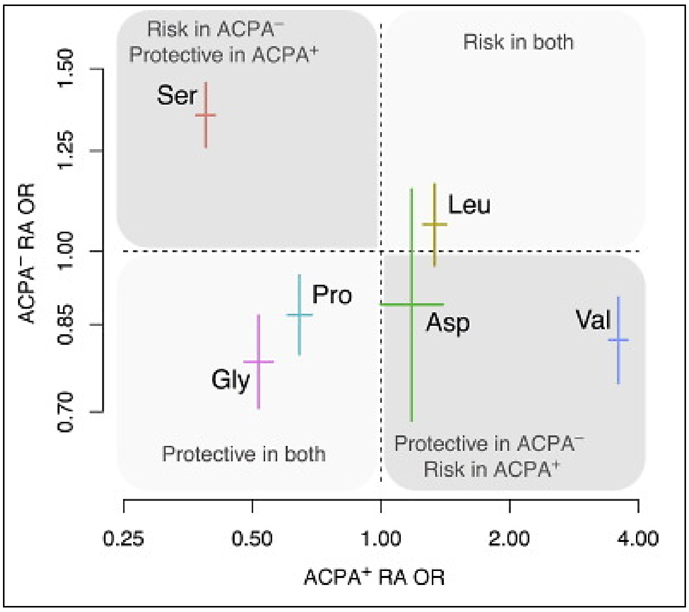


Figure 1 Distinct Effect Sizes of Amino Acid Residues at HLA-DRβ1 Position 11

Effect sizes and confidence intervals for ACPA$^+$ are shown on the x-axis and for ACPA$^-$ on the y-axis[44]. Reprinted from American journal of human genetics, 94, Han *et al.*, Fine mapping seronegative and seropositive rheumatoid arthritis to shared and distinct HLA alleles by adjusting for the effects of heterogeneity, 522-532, Copyright 2014, with permission from Elsevier.

The second pre-GWAS era association to be identified was the rs2476601, non-synonymous SNP located in the *PTPN22* gene which causes an arginine at position 620 to be replaced by tryptophan (R620W)[45]. This risk allele of rs2476601 is common in European and American populations (5-15%), although is absent to rare in African and Asian populations (0-2%) suggesting that the allele appeared late in humans in a European population[46]. The *PTPN22* R620W variant has also been associated with many other autoimmune diseases including systemic lupus erythematosus (SLE)[47], Myasthenia Gravis[48], Crohns[49], juvenile idiopathic arthritis (JIA)[50] and originally in type 1 diabetes (T1D)[51]. This suggests it has a more general autoimmune effect and as a result, several groups have studied the functional consequence of the R620W polymorphism with mixed results. Bottini *et al.*[51] and Begovich *et al.*[45] identified the association in T1D and RA respectively and studied the functional impact of the variant. Their findings showed that the 620W variant represents a gain of function allele by altering the ability of LYP, the protein encoded by *PTPN22*, to interact with Csk, a negative regulatory kinase, potentially leading to a decrease in T-cell signalling and activation. Further studies in healthy individuals and those with autoimmune diseases have corroborated these findings, showing reduced interleukin 2 (IL-2) production, decreased activity of the NFAT/AP-1 transcription factor complex, increased phosphatase activity, reduced calcium mobilisation and reduced T-cell receptor signalling[52–54]. However other studies have shown the 620W variant to be a loss of function allele[54–56], showing more efficient calcium mobilisation in T-cells, higher numbers of IL-2 producing cells and increased numbers of autoantibody producing cells in carriers of the 620W variant compared to individuals carrying the R620 variant. Overall however, most primary cell studies have found the R620W variant to have a gain of function[46]. These contrasting findings may be due to the variation having different effects on multiple pathways in the same individual or cell type or may represent disease specific effects.

Subsequent GWAS, GWAS meta-analyses and candidate gene studies prior to 2012 identified a further 33 loci associated with RA in European populations[3,10,11,38–40,57–60]. In 2012 the results of the Immunochip study[19] were published and identified an additional 14 loci to total 48 non-HLA RA associations (Figure 2). Through these genetics studies in RA, differences between ethnicities were also observed. For example, as early as 2003, Suzuki *et al.*[61] identified SNPs in the *PADI4* gene which were associated with RA in a Japanese population. This association was eventually replicated in samples of European ancestry but not robustly until 2012 by Eyre *et al.*[19]. Another example, previously mentioned, is the *PTPN22* R620W variant which, although being robustly associated with RA in European and American populations, is virtually absent in Asian populations and is not associated with RA susceptibility. A further trans-ethnic analysis of the Immunochip results combined with GWAS results and whole-genome imputation[28] identified a further 53 loci associated with RA susceptibility resulting in 101 non-HLA RA associations (Figure 3). This study also identified differences between European and Asian populations, identifying 18 variants only present in Europeans and 1 only present in Asians, although their findings did support their hypothesis that, in general, the genetic risk of RA is shared.

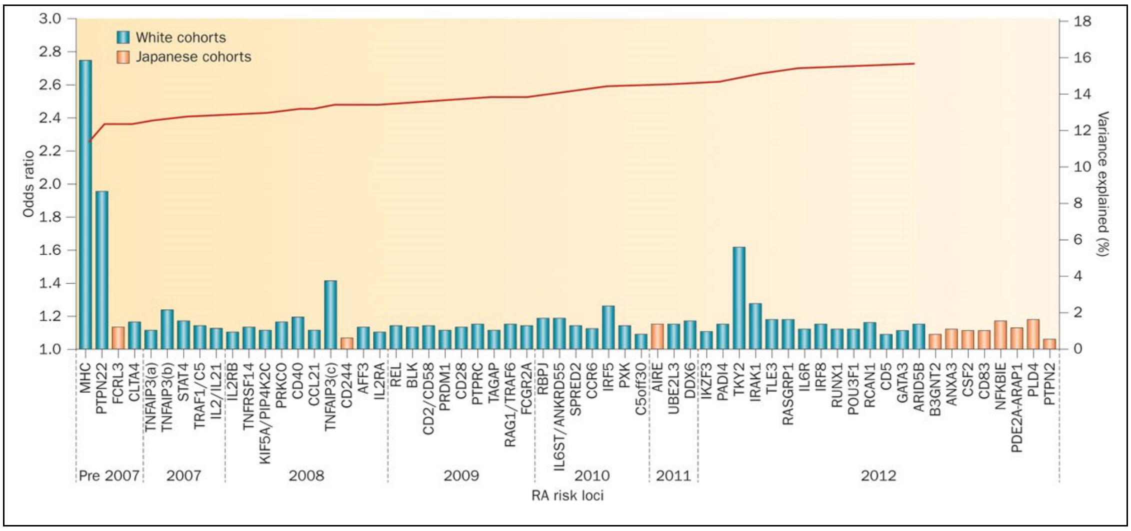


Figure 2 RA genetic susceptibility loci identified prior to 2013

Loci are shown on the x-axis and effect sizes on the y-axis. Cumulative proportion of observed variance in disease susceptibility explained is shown by the red line[62]. Reprinted by permission from Macmillan Publishers Ltd: Nature Reviews Rheumatology 9:141-153, copyright 2013.

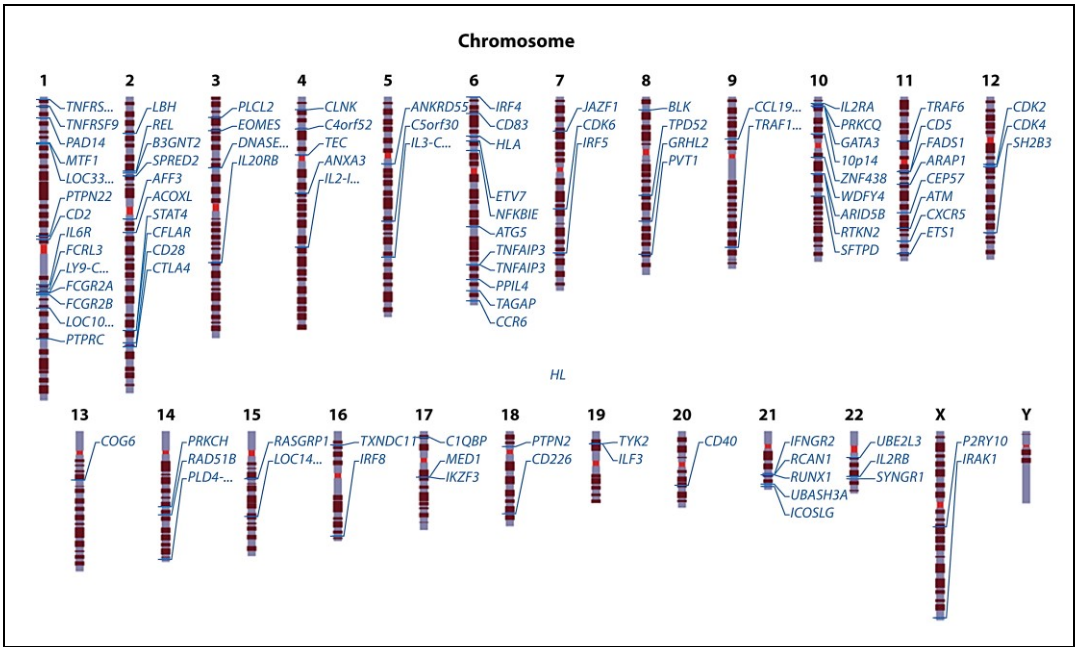


Figure 3 RA genetic susceptibility loci identified to date

Approximate chromosomal position of markers are indicated[63].

1.2.2 Shared Autoimmune Risk Loci

GWAS have therefore been successful in identifying variants and increasing our knowledge of RA genetics, implicating several loci important in disease. These include both variants which are unique to RA (~28%) and many which share associations, some substantially, with other complex diseases, primarily autoimmune disorders. For example, associations attributed to *CCL21* and *RBPJ* are currently only associated with RA (Figure 4), whereas associations attributed to *TNFAIP3*, *PTPN22*, *IL2RA* and *STAT4* are associated with multiple autoimmune diseases, such as JIA[21,50], T1D[9,51,64,65], Crohn's[66,67], ulcerative colitis (UC)[67] and SLE[47]. However, not all of the overlapping associations have the same direction of effect in all diseases. Some, such as *PTPN22*, increase an individual's risk of RA (odds ratio (OR): 1.80[28]), but are protective for Crohn's (OR: 0.79[66]). Whereas others, such as *IL6R*, are protective in RA (OR: 0.90[19]) but risk in atopic dermatitis (OR: 1.15[68]). Figure 4 shows the overlap between all genes assigned to a genetic association in RA and 12 additional autoimmune diseases from ImmunoBase. Interestingly, there is limited genetic overlap between RA and other arthritic disorders, such as JIA (Figure 4) and psoriatic arthritis (PsA)[69], compared to other unrelated autoimmune diseases, suggesting a different disease mechanism. Indeed, the most relevant cell types for RA have been epigenetically determined as CD4$^+$ T-cells and B-cells[12,70], whereas for PsA, CD8$^+$ T-cells appear to be more important in disease[69,71,72]. The non-overlapping nature of RA and PsA is also apparent from the use of therapies in disease. Although there are treatments which are used in PsA and psoriasis (Ps), which share a high degree of genetic overlap, highly effective treatments in RA, such as anti-TNF biologics, have little efficacy in PsA and Ps.

Gene	AS	ATD	CEL	CRO	JIA	MS	PBC	Ps	RA	SLE	T1D	UC	OD
ABHD6									■	■			■
ACOXL									■		■		■
AFF3			■		■	■			■		■		
ANKRD55			■	■	■	■			■		■		■
ARID5B									■	■			
ATG5				■					■	■		■	■
ATM									■				
BACH2	■	■	■	■		■			■		■		■
BLK									■	■			■
C1QBP									■				
C5									■				
C5orf30								■	■				
CASP8									■				
CCL19									■				
CCL21									■				
CCR6		■		■					■			■	■
CD226				■		■			■		■	■	■
CD28		■	■						■		■		■
CD40				■		■			■			■	■
CD5				■		■			■			■	■
CDK2								■	■		■		■
CDK6									■				
CEP57									■				
CFLAR									■				
CLNK									■				■
COG6					■				■				
CSF3				■		■	■		■	■	■	■	■

CTLA4		■	■						■		■		■
CXCR5			■	■		■	■		■			■	■
DDX6			■	■		■	■		■				■
DNASE1L3									■	■			■
ELMO1			■			■		■	■		■		
EOMES						■			■				
ETS1			■					■	■	■			
ETV7						■			■				
FADS1				■					■			■	■
FADS2				■					■			■	■
FADS3				■					■				■
GATA3									■				
GRHL2									■				
HLA-DRB1	■	■	■	■	■	■	■	■	■	■	■	■	■
IKZF3				■		■	■		■	■	■	■	■
IL2			■	■	■				■		■	■	■
IL20RB									■				
IL21			■	■	■				■		■	■	■
IL2RA		■		■	■	■			■		■	■	■
IL6R	■				■				■		■		
IL6ST				■					■			■	■
ILF3						■		■	■				
INPP5B									■				
IRAK1			■						■	■			
IRF4			■					■	■		■		
IRF5				■			■		■	■		■	■
IRF8				■		■	■		■			■	■
JAZF1				■	■	■			■	■		■	■
LBH									■				
LINC01343									■				
LOC100506023									■	■			■
LOC100506403					■				■				
LOC145837									■				
MED1				■		■	■		■	■	■	■	■
MMEL1		■	■	■		■			■			■	■
MTF1									■				
NFKBIE									■				
P2RY10									■				
PADI4									■				
PLCL2						■	■		■				
POU3F1									■				
PPIL4									■				
PRKCH									■		■		
PRKCQ			■						■		■		
PTPN11	■		■		■		■		■	■	■		■
PTPN2			■	■	■				■		■	■	■
PTPN22		■		■	■				■	■	■	■	■
PVT1						■			■				■
PXK									■	■			■
RAD51B			■	■	■	■	■		■	■	■		■
RASGRP1				■					■	■	■	■	■
RBPJ									■				
RCAN1									■				

REL			■	■		■		■	■			■	■
RTKN2									■	■			
RUNX1					■				■				
SH2B3	■		■		■		■		■	■	■		■
SMIM20									■				
SPRED2			■	■		■			■		■	■	■
STAT4			■	■	■	■	■		■	■	■	■	■
SYNGR1				■			■		■			■	■
TAGAP			■	■		■		■	■		■	■	■
TLE3									■				
TNFAIP3			■	■		■	■	■	■	■	■	■	■
TNFRSF14		■	■	■		■			■			■	■
TPD52									■				
TRAF1									■				
TXNDC11									■				
TYK2	■		■	■	■	■	■	■	■	■	■	■	■
UBASH3A			■						■		■		■
UBE2L3			■	■	■	■		■	■	■		■	■
WDFY4									■	■			
YDJC			■	■	■	■		■	■	■		■	■
ZNF438						■			■				

Figure 4 Comparison of RA associated genes against 12 additional autoimmune diseases taken from ImmunoBase (http://www.immunobase.org)

Disease abbreviations are as follows: AS – Ankylosing spondylitis; ATD – Autoimmune thyroid disease; CEL – Coeliac disease; CRO – Crohn's disease; JIA – Juvenile idiopathic arthritis; MS – Multiple sclerosis; PBC – Primary biliary cirrhosis; Ps – Psoriasis; RA – Rheumatoid arthritis; SLE – Systemic lupus erythematosus; T1D – Type 1 diabetes; UC – Ulcerative colitis; OD – Other diseases.

1.2.3 RA Clinical Subtypes

GWAS have also highlighted genetic differences between subtypes of RA. As mentioned previously, RA can be broadly classified into two subtypes, ACPA$^+$ and ACPA$^-$, based on the presence of antibodies against cyclic citrullinated peptides (anti-CCP). These subtypes are clinically indistinguishable at diagnosis, but the presence of anti-CCP antibodies predicts disease severity and radiological damage[73,74], with ACPA$^+$ RA patients having a more severe disease. As such, ACPA$^+$ RA patients are seen more often at rheumatology clinics and recruited onto genetics studies and therefore the majority of RA genetics studies have been performed on the ACPA$^+$ subtype. There is also thought that these subtypes are genetically different and may in fact represent two clinically different conditions.

Initially, based on twin studies, ACPA$^-$ RA was estimated to have the same heritability as ACPA$^+$ RA (~60%)[75], but these estimates have since been revised to 50% and 20% for ACPA$^+$ and ACPA$^-$ RA respectively[76]. Despite this reduction in heritability in ACPA$^-$ disease, the effect of HLA is much lower than in ACPA$^+$ patients and therefore additional ACPA$^-$ RA genetic associations are likely to exist[77]. However, both GWAS and candidate gene studies have had little success in identifying variants associated with ACPA$^-$ RA and in addition, many have only been reported in single studies without independent replication. The first comprehensive analysis of ACPA$^-$ RA, based on the GWAS meta-analysis by Stahl *et al.*, identified 6 loci, already associated in ACPA$^+$ RA, which are also associated with ACPA$^-$ RA (Table 1)[78].

Table 1 Schematic classification of RA susceptibility loci into three categories depending on their association pattern in anti-CCP positive and negative RA

Category	Associations	Locus Name
1	Both CCP positive and negative RA, stronger in CCP positive RA	*PTPN22*, *TNFAIP3*[a]
2	Both CCP positive and negative RA, equally strong in both	*ANKRD55*, *BLK*, *C5orf30*, *STAT4*
3	CCP positive RA only, significant difference between CCP positive and negative RA	*AFF3*, *CCR6*, *CCL21*, *IL2RA*, *CD28*, *CD40*, *PXK*, *REL*, *RBPJ*, *TNFRSF14*, *TNFAIP3*[b]
Not classifiable	CCP positive RA only, but no significant difference between CCP positive and negative RA	All others

[a] rs6920220; [b] rs5029937. Reproduced from Genetic markers of rheumatoid arthritis susceptibility in anti-citrullinated peptide antibody negative patients, Viatte *et al.*[78], 71,1984-90, 2012 with permission from BMJ Publishing Group Ltd.

A recent study by Viatte *et al.*[77] utilising ACPA$^-$ RA data generated using the Immunochip array, supplemented by 1,044 replication samples, confirmed existing loci and identified two novel ACPA$^-$ specific loci (*PRL* & *NFIA*). Together with other confirmed or suggestive loci from other studies, this results in 14 ACPA$^-$ RA associated loci. Importantly, this study concluded that, given its sample size, if ACPA$^-$ RA had similar genetic architecture to that of ACPA$^+$ RA, it would have been equivalent to a study conducted in 2012 (for example Eyre *et al.*[19], 48 non-HLA loci). However, it is clear that the effect sizes of the ACPA$^-$ associations are smaller than ACPA$^+$ RA. These findings could be explained by the smaller genetic contribution to ACPA$^-$ RA but could also suggest that the ACPA$^-$ RA subtype is itself a heterogeneous sample population. Despite this, it is clear that ACPA$^+$ and ACPA$^-$ RA subtypes are genetically different subsets which only share partial genetic overlap.

1.2.4 Missing Heritability

Despite the success of GWAS and candidate gene studies in RA, most of the associations identified have modest effect sizes (OR <1.5) (Figure 2) and altogether only account for ~19.5% of the total heritability for RA[28]. This outcome is also true for many other complex diseases and has been termed the ‘missing heritability’ of a disease. The missing heritability could be due to many factors, including more, as yet undiscovered, associated variants, rarer and structural variants which are under investigated, the multiplicative effect of having a burden of risk variants and the unknown/under investigated interactions between genetics and the environment. However, there is conflicting opinion regarding the concept and cause of ‘missing heritability’.

1.2.5 T-cells and RA

The role of T-cells was implicated in RA pathogenesis many years ago, mostly due to the association with the HLA region[79]. HLA proteins are responsible for presenting short (<30 amino acids) foreign peptides or antigens to T-cells, the first step in the process of T-cell activation, as part of the adaptive immune response. Additionally, T-cells isolated from RA synovial tissue show increased expression of markers of antigen exposure, CD45RO and CD27, relative to circulating T-cells[79].

Further genetic evidence since 2005 has added to this hypothesis. For example, *PTPN22*, an early discovered risk loci, is responsible for inhibition of T-cell activation by restricting signalling downstream of the T-cell receptor (TCR)[80] and more recently Maine *et al.*[81] and Brownlie *et al.*[82] demonstrated a link between *PTPN22* and the development of regulatory T-cells (Tregs) in the thymus. Additional associations with *CCR6*, a chemokine receptor, expressed by CD4$^+$ type 17 T helper (T$_H$17) cells and associated with interleukin 17 (IL-17) RA sera levels[83], and *IL2RA*, correlated with mRNA and surface protein levels in CD4$^+$ naïve and memory T-cells[84], add further support for the role of T-cells.

Pathway analysis of RA risk loci also support the role of T-cells, highlighting immune pathways such as T-cell activation and differentiation, antigen processing and presentation and JAK/STAT signalling[85]. Indeed, many genes involved in signalling between dendritic cells and T-cells reside in RA associated regions (Figure 5). Studies investigating enrichment in gene expression data, DNA methylation and other epigenetic marks have identified RA genetic associations to be enriched in T-cells in general[86,87] and specific enrichment has been found in T$_H$17[12], CD4$^+$ regulatory T cells[87] and CD4$^+$ effector memory T-cells[88].

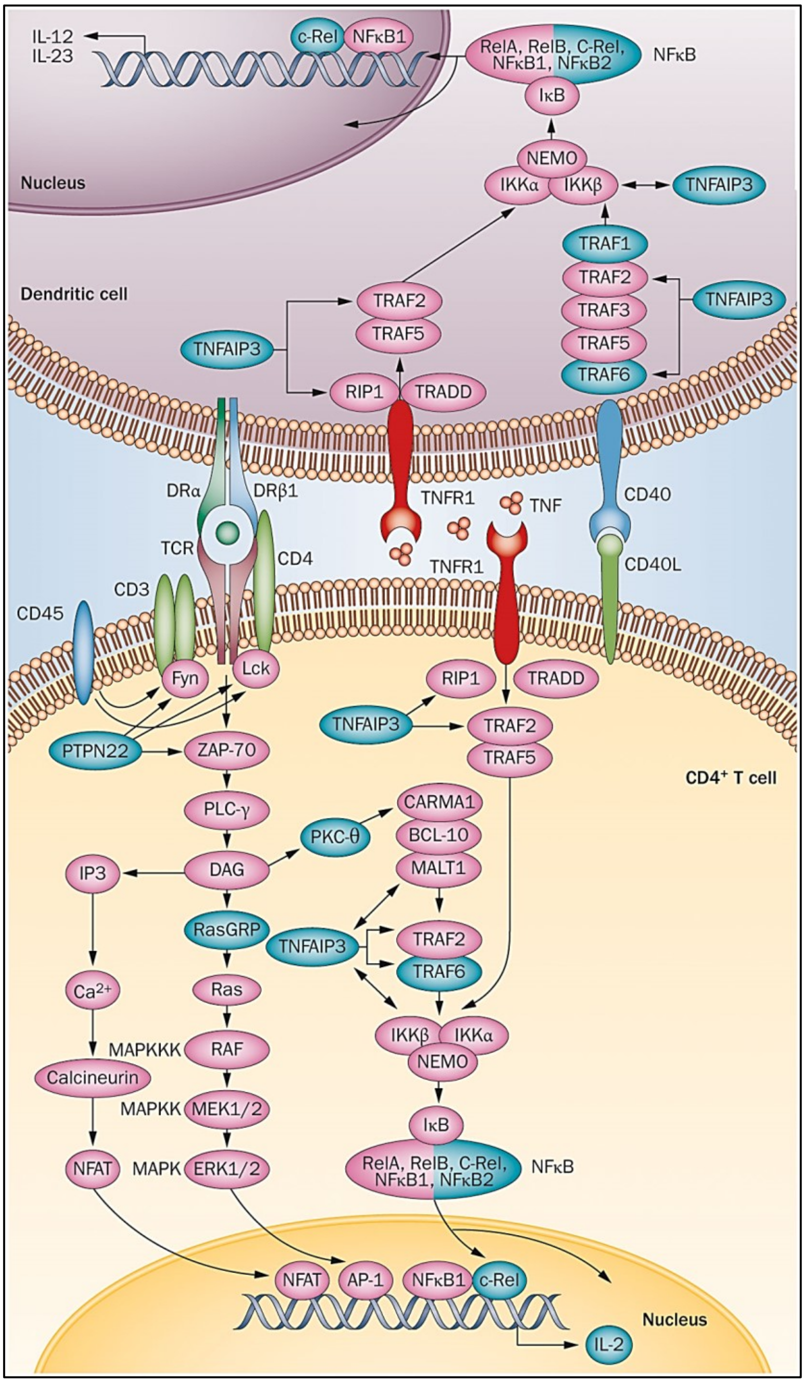


Figure 5 T-cell–dendritic-cell dialogue

Genes filled in blue are encoded by genes within RA susceptibility loci[62]. Reprinted by permission from Macmillan Publishers Ltd: Nature Reviews Rheumatology 9:141-153, copyright 2013.

Finally, evidence comes from a current biologic therapy, abatacept, used to treat RA. Abatacept is an immunoglobulin fusion protein based on cytotoxic T-lymphocyte protein 4 (CTLA-4). The CTLA-4 protein (also known as CD152) is expressed on the surface of T-cells and is an important negative regulator of T-cell activation[89]. Importantly, variants in the *CTLA4* gene region are also associated with RA[19,28,38].

This evidence clearly shows the importance of T-cells in RA pathogenesis and the genetic evidence suggests that changes in T-cells in RA patients are likely to be a cause, rather than a consequence, of disease. However, studies have also suggested a role for B-cells, part of the adaptive immune system and responsible for the secretion of antibodies. B-cells can also present antigens and secrete cytokines, immune signalling molecules, and can be activated by T-cells[90]. Studies have shown a therapeutic benefit for B-cell depletion[91,92] and the use of a B-cell biologic therapy, rituximab, targeting CD20 expressed on the surface of B-cells[93] supports this. Additional evidence implicating B-cell signalling pathways[28], enrichment of B-cell specific enhancers[12] and genes involved in B-cell function[63], adds further support for this hypothesis. These findings highlight the complexity of RA pathogenesis and the interplay between different cells of the immune system.

1.2.6 Drug Targets in RA

Current treatment options for RA do not always prove effective, they can cause unacceptable side-effects (adverse events) or just simply not control the disease sufficiently (inefficacy). Traditional disease-modifying anti-rheumatic drugs (DMARDs), such as methotrexate and sulfasalazine, are the first step in the treatment of severe RA, however, up to two thirds of patients fail to respond, either due to adverse events or inefficacy[94]. Therapies based upon biological proteins, termed biologics, were introduced in the late nineties and are new type of DMARD which, despite being expensive, have proven to be effective in the treatment of RA[95]. They target specific molecules involved in the immune response, such as tumour necrosis factor (TNF), interleukin 6 receptor (IL-6R) and CTLA-4, to supress the immune response and as a result reduce disease activity. Although there are examples of biologic therapies which are based on RA associated genes, such as abatacept (*CTLA4*) or which are antagonists of RA associated genes, such as tocilizumab (*IL6R*) as well as pathways identified by RA genetics, such as tofacitinib (janus kinase (JAK) inhibitor), etanercept (anti-TNF) and rituximab (B-cell surface molecules (CD20)), none of the therapies currently used to treat RA were developed based on RA genetics.

Okada *et al.* evaluated the potential of drug discovery in RA by testing if any genes identified either as RA risk genes or by a direct protein-protein interaction (PPI) network were targets of existing RA drugs[28]. They found that 27 targets for approved RA drugs showed significant overlap with 98 RA risk genes and 2,332 PPI genes. The authors therefore concluded that as genetics was successful in identifying RA drug targets, it also has the potential to be useful in drug target validation. Further work by our group has identified 41 targets for 106 existing drugs from work informed by genetics for RA[96].

1.2.7 Future of RA Genetics

Although RA GWAS has therefore led to major strides in understanding disease, these findings are based on a relatively imprecise knowledge of the exact genes, cell types and pathways implicated in disease. Few have been functionally explored and associated regions have been labelled with the closest, most compelling candidate with little or no evidence to support their candidature. Many show no obvious role in RA pathogenesis as they lie in non-coding regions. In fact, of the 101 loci identified by Okada *et al.*[28], whilst 50% have expression quantitative trait loci (eQTL) data or are non-synonymous variants, their involvement in RA pathogenesis is unclear. The next challenge for RA is to functionally determine the effect of these variations and identify or confirm their target genes to fully explore RA disease susceptibility. This then has the potential to provide novel, effective therapies, thus decreasing the economic burden of RA and improve the quality of life for RA patients.

1.3 Functional Genomics and the post-GWAS era

At a similar time to the GWAS era, a large scale project to characterise the functional elements of the human genome was initiated. A pilot study by the ENCyclopaedia Of DNA Elements (ENCODE) international consortium on 1% of the genome was increased to the whole genome in 2007. This project studied several functional elements such as transcription factor binding sites, DNase I hypersensitivity and histone modifications across multiple cell lines using newly developed next-generation sequencing (NGS) based experimental methods (Figure 6).

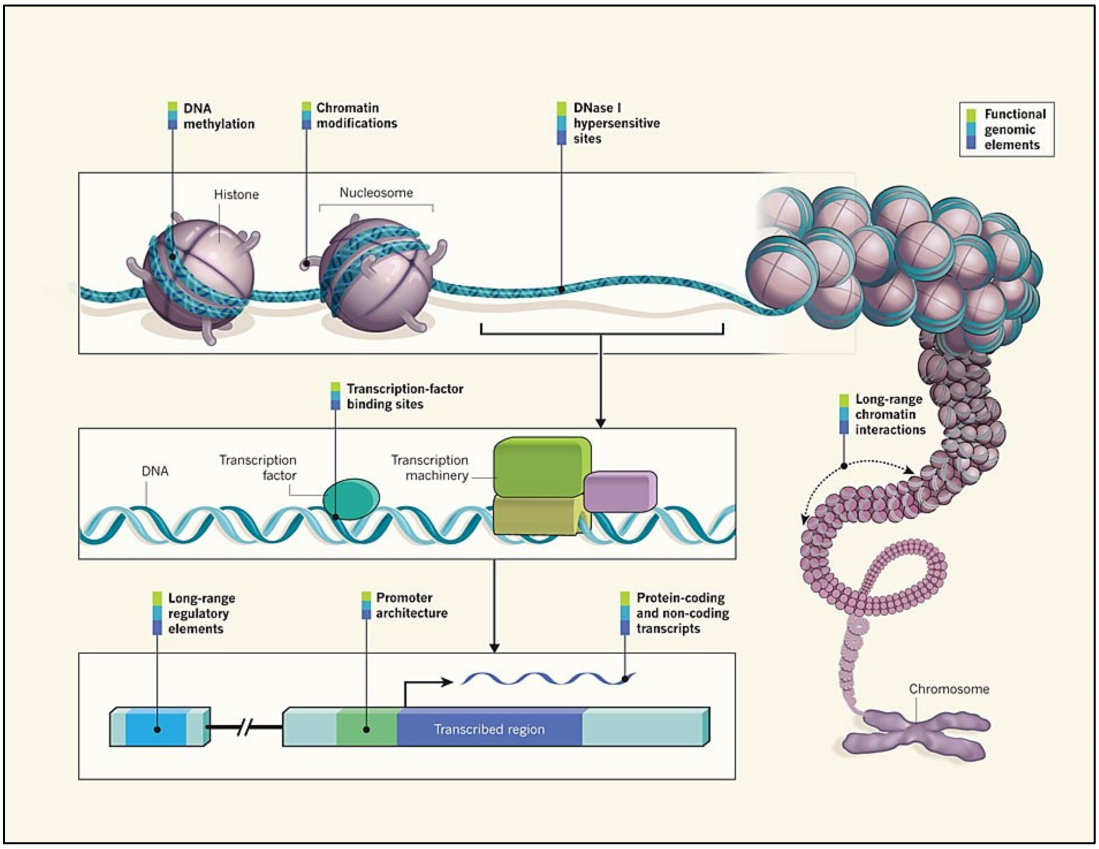

DNA methylation
Chromatin modifications
DNase I hypersensitive sites
Functional genomic elements
Histone
Nucleosome
Transcription-factor binding sites
Long-range chromatin interactions
DNA
Transcription factor
Transcription machinery
Long-range regulatory elements
Promoter architecture
Protein-coding and non-coding transcripts
Transcribed region
Chromosome

Figure 6 Overview of the ENCODE project

Experimental approaches are indicated at the appropriate resolution[97]. Reprinted by permission from Macmillan Publishers Ltd: Nature 489:52-55, copyright 2012.

DNA is often thought of as a two dimensional linear string however within the cell nucleus it is heavily condensed into chromatin, a DNA-protein complex, comprising of DNA wrapped around proteins called histones to produce nucleosomes, approximately 11nm in diameter. These nucleosomes are further compressed, folded and coiled to compact them enough to fit in the nucleus of the cell (Figure 7). However, histones can be modified to change how tightly packed that region of the genome is and observations of these modifications can indicate how active the region is and therefore how likely it is to be involved in gene regulation. Similarly DNase I hypersensitivity data can be used to tell how open or accessible the region is to other gene regulators such as transcription factors. Collectively, these observations constitute the cells epigenome and allow researchers to characterise regions of the genome into functional classes and determine their relevance. For example, correlating this data with all SNPs in high LD with the GWAS identified SNP allows genetics researchers to identify which of these potential candidate SNPs is most likely to be causal and therefore prioritise these for expensive functional follow-up studies.

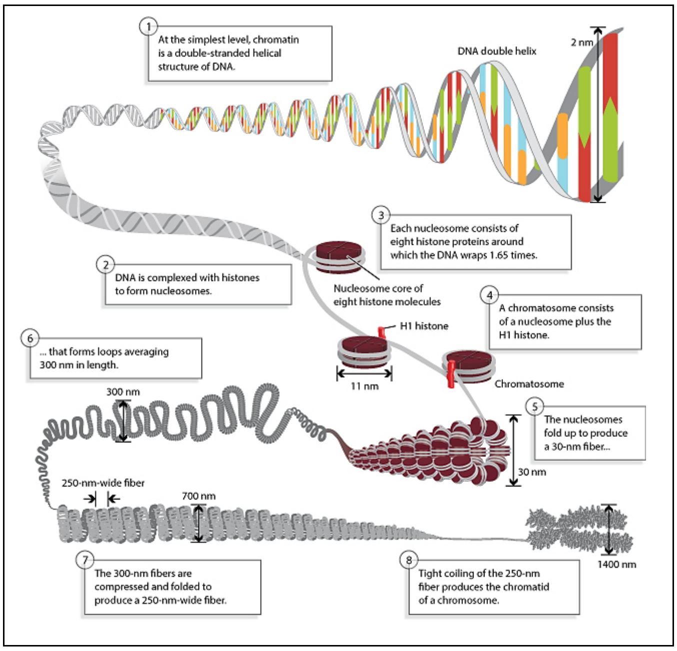

1
At the simplest level, chromatin is a double-stranded helical structure of DNA.
DNA double helix
2 nm
3
Each nucleosome consists of eight histone proteins around which the DNA wraps 1.65 times.
2
DNA is complexed with histones to form nucleosomes.
Nucleosome core of eight histone molecules
4
A chromatosome consists of a nucleosome plus the H1 histone.
H1 histone
6
... that forms loops averaging 300 nm in length.
11 nm
Chromatosome
300 nm
5
The nucleosomes fold up to produce a 30-nm fiber...
30 nm
250-nm-wide fiber
700 nm
1400 nm
7
The 300-nm fibers are compressed and folded to produce a 250-nm-wide fiber.
8
Tight coiling of the 250-nm fiber produces the chromatid of a chromosome.

Figure 7 Compaction of DNA into chromatin

Numbered boxes indicate compaction method at each stage[98].

Since the ENCODE project, complementary large whole-genome epigenomics projects have been initiated. The NIH Roadmap Epigenomics Mapping Consortium started the Roadmap Epigenomics Project with the aim to characterise the epigenomes of primary and *ex vivo* tissues used to represent normal human tissues involved in disease but do not target other non-epigenetic transcriptional regulators such as transcription factors. Similar projects also include the Blueprint epigenome project (http://www.blueprint-epigenome.eu/index.cfm) and the International Human Epigenome Consortium (IHEC), including the Canadian Epigenetics, Environment and Health Research Consortium (CEEHRC) Epigenomic Platform Program (http://www.epigenomes.ca/), the Deutsches Epigenom Programm (DEEP) (http://www.deutsches-epigenom-programm.de/) and the Core Research for Evolutional Science and Technology (CREST) (http://crest-ihec.jp/english/index.html) projects which focus on different cell types or experimental aims.

It is clear that utilising this data in addition to the generation of new disease focused experiments will be essential to fully translate genetic findings to progress the understanding of the genetic basis of complex disease. Post-GWAS bioinformatics will be fundamental to this process to analyse, exploit and integrate these large cross-disciplinary datasets and explore disease aetiology and further the discovery of new treatment options.

2 Identification of Causal SNPs and Their Function (Publications 1 and 2)

2.1 Background

As GWAS identified increasing numbers of SNPs for many complex diseases, it became apparent that the vast majority of these variants were located outside traditional protein coding regions of the genome and therefore predicted to have a role in gene regulation. Techniques to study gene regulation were already established but without specific hypotheses would result in expensive and time-consuming experiments that may not identify any effect on disease susceptibility. It would therefore be necessary to narrow down the number of potential SNPs and formulate specific hypotheses to test and prioritise these for future work.

The production phase ENCODE project[30,99] was initiated in 2007 to study the functional elements of the whole human genome and has successfully generated data on multiple cell lines using many different unique and complementary experimental techniques (Figure 6). This data is publicly available and can be accessed through the Univeristy of California, Santa Cruz (UCSC) Genome Browser[100]. Different combinations of DNase I hypersensitivity (HS) sites and histone marks are indicative of certain 'chromatin states', for example, active enhancers or promoters (Figure 8). By aggregating this data for any given SNP or region, researchers can infer the chromatin state and build up evidence to either strengthen or weaken the case towards the likelihood for a SNP being functionally relevant. For example, if an associated variant lies in an area demonstrating DNase I HS and active promoter histone marks (H3K4me3 and H3K27ac), this shows that this region is 'open' and accessible to other regulatory or transcriptional proteins and is an active promoter, which would support the functional role of this variant. Conversely, if the region lacked DNase I HS data and active promoter histone marks, it would suggest that the region is 'closed' and transcriptionally inactive. Additionally, if there was evidence of transcription factor or PolII binding, this would strengthen the case further still.

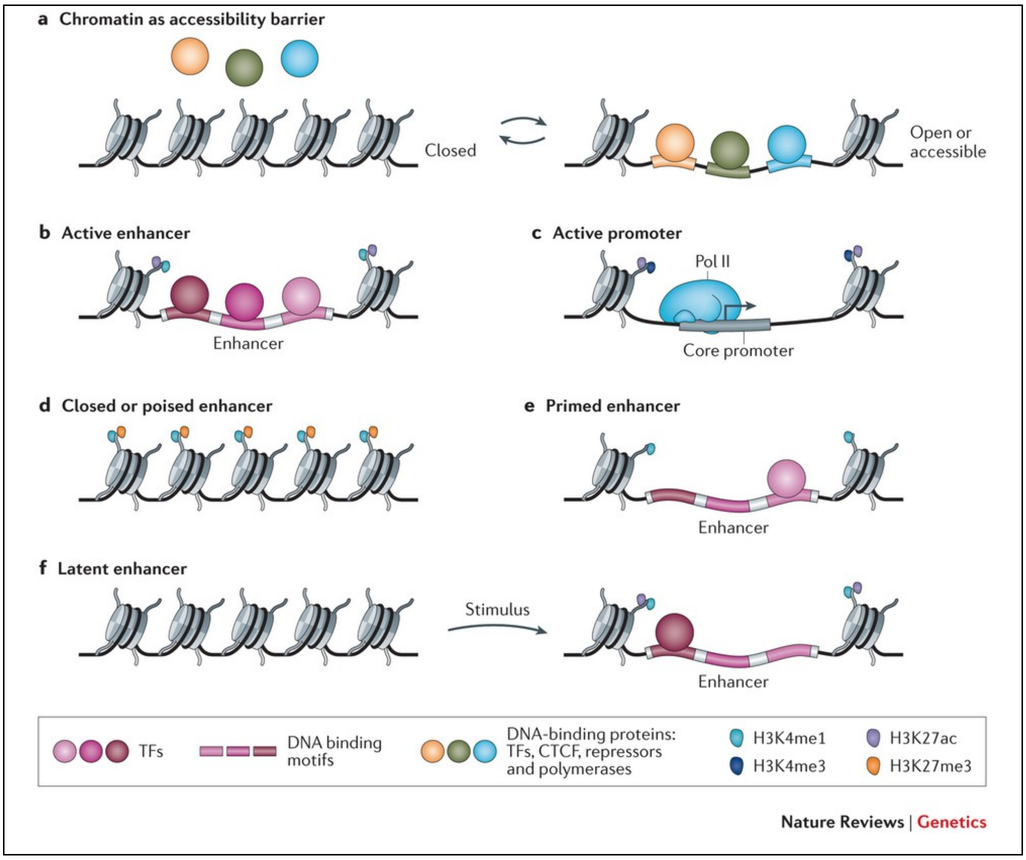


Figure 8 Various chromatin states

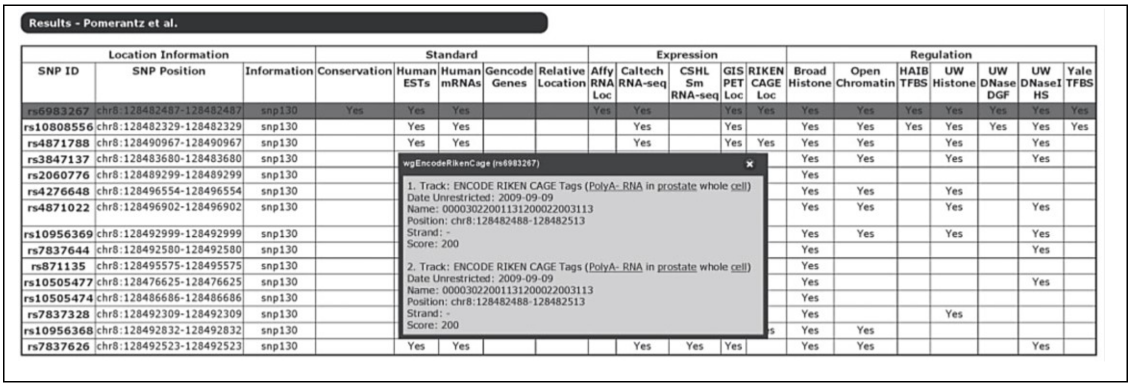
Each chromatin state is characterised by DNase I HS and histone marks according to the key[101]. Reprinted by permission from Macmillan Publishers Ltd: Nature Reviews Genetics 15:272-286, copyright 2014.

However, while this resource has proved to be invaluable, the sheer wealth of information available has also provided a challenge to many as fully mining and utilising this data can be prohibitive and researchers often lack the appropriate skills to identify and aggregate information across the various experiments and cell types.

2.2 Aims and achievements

The aim of this work was therefore to provide researchers with an easy to use tool that could automatically interrogate, assimilate and aggregate this data for selected SNPs and present it in the most efficient way so that researchers could identify the most likely causal SNPs amongst the potential candidates, allowing them to evaluate and interpret the data at a high level but still have access to the complete underlying data.

To address this I developed ASSIMILATOR[29] to quickly and effectively query the UCSC database and present the results in a user friendly manner (Figure 9). ASSIMILATOR was written in Perl and directly queries the UCSC database to interrogate the ENCODE data. The ENCODE data is stored as genomic features (chromosome, start and end co-ordinates) across multiple tables organised by cell type and experimental features, such as DNase I HS, H3K4me1 and CTCF transcription factor and due to the amount of data available it is non-trivial to integrate all the evidence and interpret results. ASSIMILATOR therefore summarises each broad experiment, such as DNase I HS, histone marks and transcription factor binding into a simple table showing presence or absence of experimental features. However, access to the full underlying data is still possible by 'clicking' the relevant experimental summary. This displays all of the overlapping features for that experiment for the queried SNP within the main page (Figure 9), allowing the user to easily compare multiple experiments or query SNPs. Additionally the user is able to link to the UCSC browser to easily visualise the experimental features without having to manually select, check and add relevant tracks.



Results - Pomerantz et al.

Location Information			Standard					Expression					Regulation						
SNP ID	SNP Position	Information	Conservation	Human ESTs	Human mRNAs	Gencode Genes	Relative Location	Affy RNA Loc	Caltech RNA-seq	CSHL Sm RNA-seq	GIS PET Loc	RIKEN CAGE Loc	Broad Histone	Open Chromatin	HAIB TFBS	UW Histone	UW DNase DGF	UW DNaseI HS	Yale TFBS
rs6983267	chr8:128482487-128482487	snp130	Yes	Yes	Yes			Yes	Yes		Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
rs10808556	chr8:128482329-128482329	snp130		Yes	Yes				Yes		Yes		Yes	Yes	Yes	Yes	Yes	Yes	Yes
rs4871788	chr8:128490967-128490967	snp130		Yes	Yes				Yes		Yes	Yes	Yes	Yes		Yes		Yes	
rs3847137	chr8:128483680-128483680	snp130											Yes	Yes		Yes		Yes	
rs2060776	chr8:128489299-128489299	snp130											Yes						
rs4276648	chr8:128496554-128496554	snp130											Yes	Yes		Yes			
rs4871022	chr8:128496902-128496902	snp130											Yes	Yes		Yes		Yes	
rs10956369	chr8:128492999-128492999	snp130											Yes	Yes		Yes		Yes	
rs7837644	chr8:128492580-128492580	snp130											Yes					Yes	
rs871135	chr8:128495575-128495575	snp130											Yes						
rs10505477	chr8:128476625-128476625	snp130											Yes					Yes	
rs10505474	chr8:128486686-128486686	snp130											Yes						
rs7837328	chr8:128492309-128492309	snp130											Yes			Yes			
rs10956368	chr8:128492832-128492832	snp130										s	Yes	Yes					
rs7837626	chr8:128492523-128492523	snp130		Yes	Yes				Yes	Yes	Yes		Yes	Yes				Yes	

wgEncodeRikenCage (rs6983267)

1. Track: ENCODE RIKEN CAGE Tags (PolyA- RNA in prostate whole cell)
Date Unrestricted: 2009-09-09
Name: 00003022001131200022003113
Position: chr8:128482488-128482513
Strand: -
Score: 200

2. Track: ENCODE RIKEN CAGE Tags (PolyA- RNA in prostate whole cell)
Date Unrestricted: 2009-09-09
Name: 00003022001131200022003113
Position: chr8:128482488-128482513
Strand: -
Score: 200

Figure 9 Example of ASSIMILATOR output

The results are shown for Pomerantz *et al.*[102] with the causal SNP highlighted.

However, allowing access to all experimental data can take a large amount of time, due to the complexity and amount of data held in the UCSC database. Therefore several systems have been implemented to speed up ASSIMILATOR and simplify its use. The first of these stores existing track information in an extensible markup language (XML) file, which is automatically updated with new tracks, reducing the number of database queries needed to obtain information about each track. The next speed improvement utilises multi-core processing to allow ASSIMILATOR to query multiple SNPs simultaneously, reducing execution time by up to eight times over a single core implementation. The final improvement enhances usability, providing a unique token which allows users to submit a SNP query and return at a later date to retrieve the results from a MySQL™ database. The results are presented as an overview in a web page which can be viewed in a standard web browser. This summary page can then be used to access the underlying data by selecting the relevant experiment and SNP.

2.3 Contribution to the literature

ASSIMILATOR was the first tool to be developed which aggregated data from ENCODE to annotate a list of regions and allowed researchers to easily identify and prioritise potential causal candidate variants for further investigation. This tool has been used to annotate other associated variants[103,104] and this is exemplified in the work by Orozco *et al.*[31] describing a large extended GWAS in RA, in which this tool was used to assess the functional impact of variants associated at the 22q12 locus. Evidence was discovered suggesting an associated variant, rs1043099, and correlated variants map to sites of transcription factor binding and open chromatin. Coupled with histone modification evidence, this suggests that these associated variants could affect gene regulation.

Shortly following the publication of ASSIMILATOR, two groups from the Broad Institute and the Center for Genomics and Personalized Medicine at Stanford University have developed similar tools, HaploReg[105] and RegulomeDB[106], augmenting them with different features. Unlike ASSIMILATOR which includes a self-updating procedure and queries the ENCODE data held at UCSC each time allowing retrieval of the most up-to-date information, both HaploReg and RegulomeDB rely on locally hosted database snapshots. As such they are faster than ASSIMILATOR but require manual bulk updates to integrate newly released data. Both tools have additional features and have been updated and developed since their initial release.

HaploReg supplemented the ENCODE data with transcription factor (TF) position weight matrices (PWMs) from TRANSFAC[107] and JASPAR[108] to annotate variants by their effect on protein binding. Additionally, all SNPs in LD with the query SNPs were included by utilising data from the 1000 genomes project[109] and storing it locally. This has both advantages as it improves user friendliness and additional analysis steps but removes an element of control from the user and the ability to use custom LD panels, for example from a disease reference panel, or specific locations, such as new possibly rare SNPs. To overcome the first limitation HaploReg allows users to disable the LD selection and enter multiple query SNPs to test. Subsequent releases of HaploReg added data from the epigenomics roadmap project[110], eQTLs from the GTEx project[111], an updated SNP database and expanded PWM data.

RegulomeDB also builds on the ENCODE data by adding PWMs and additional annotations from ChIP-Seq, eQTL, DNase I sensitivity QTLs (dsQTLs) and ChIP-exo experiments. RegulomeDB is unique as, in addition to the underlying data, it also provides the user with a score rating how likely the query SNP is to have a functional consequence. This score is however based on the presence of certain features, rather than a statistical measure, using the likelihood of observing the individual experimental features and aggregating them into measure of significance. For example, to attain a score of 1a, the SNP must show evidence of an eQTL, TF binding, matched TF motif, matched DNase I footprint and DNase I peak. Since the initial release of RegulomeDB the database has been updated to the 2012 ENCODE data freeze and additional data on chromatin states from the epigenomics roadmap project, DNase footprinting, PWMs, and DNA methylation has also been added.

More recently, tools have been released which utilise statistical methods to score variants using various annotation sources or test for enrichment within a specific set of annotations, for example, histone marks[70,112,113]. These do not utilise full annotation datasets or test all SNPs and do not provide information on individual variants but on disease associations in general and therefore are less utilised for SNP prioritisation than HaploReg and RegulomeDB.

Further tools have been developed which score variants based not only on functional annotation but evolutionary fitness[114] or deleteriousness[115]. Kircher *et al.* have developed a framework, combined annotation-dependent depletion (CADD), which compares the annotations of 'fixed' derived alleles with simulated variants[115]. CADD is based on the assumption that deleterious mutations are removed, or depleted, by natural selection in fixed variation, but not in simulated variation. Annotations are obtained from various sources, such as the Ensembl Variant Effect Predictor[116] (VEP), ENCODE and the UCSC genome browser and a matrix of 29.4 million fixed and simulated variants (50:50 ratio) against 63 annotations is produced. Scores are then precomputed by applying the average of ten models trained on the labelled matrix. This method performs particularly well for nonsense variants, however less so for non-coding variants, producing much lower scores compared to nonsense variants (Figure 10) and as such, its application to GWAS SNPs may be limited. This limitation has also been commented on by Gulko *et al.* who later released fitCons[114], although their comparison was later refuted by Cooper *et al*[117].

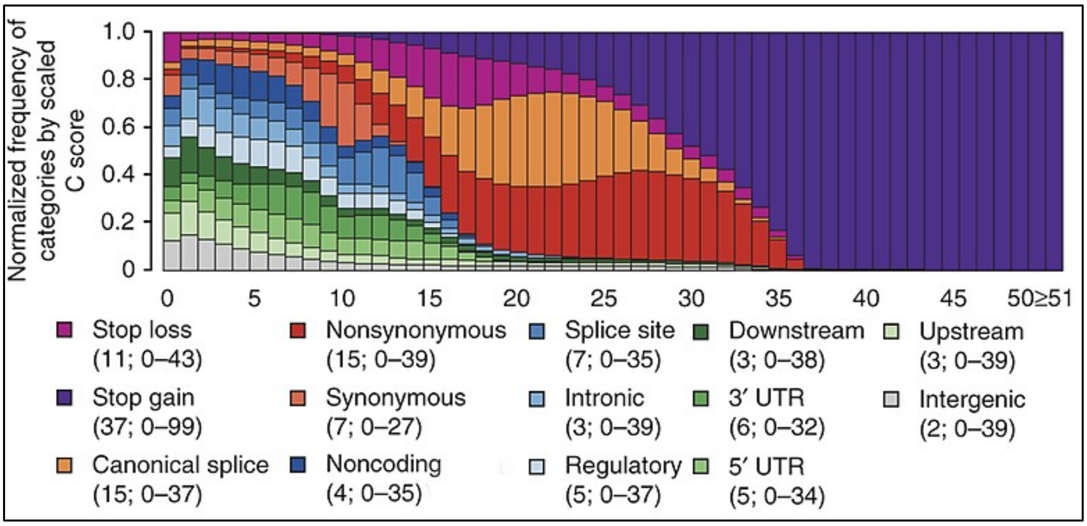


Figure 10 Relationship of scaled C scores and categorical variant consequences. Proportion of substitutions with a specific consequence after first normalizing by the total number of variants observed in that category. The legend includes in parentheses the median and range of scaled C score values for each category. Consequences were obtained from Ensembl VEP. Adapted by permission from Macmillan Publishers Ltd: Nature Genetics 46:310-315, copyright 2014.

FitCons[114], developed by Gulko *et al.*, seeks to assign a probability that a variation will affect fitness to each position in the genome. This score can then be used as an evolution-based measure of potential function. To calculate the score, functional annotation data from ENCODE, primarily DNase I, RNA-Seq and histone ChIP-Seq, is clustered to produce 624 distinct functional genomic classes. This task is simplified by using chromatin states, as opposed to full ChIP-Seq data, defined by ChromHMM, discussed later. This is followed by estimating the fraction of sites under selection, using the INSIGHT method, by functional class and assigning this score to each position belonging to that class. Similar to CADD, fitCons assigns the highest scores to coding variants and the lowest score to intergenic variants showing little or no evidence for functional enrichment (Figure 11). Interestingly, the authors also show that the performance of fitCons for non-coding variants outperforms other methods (Figure 12). However, the method appears to have been less widely adopted compared to CADD.

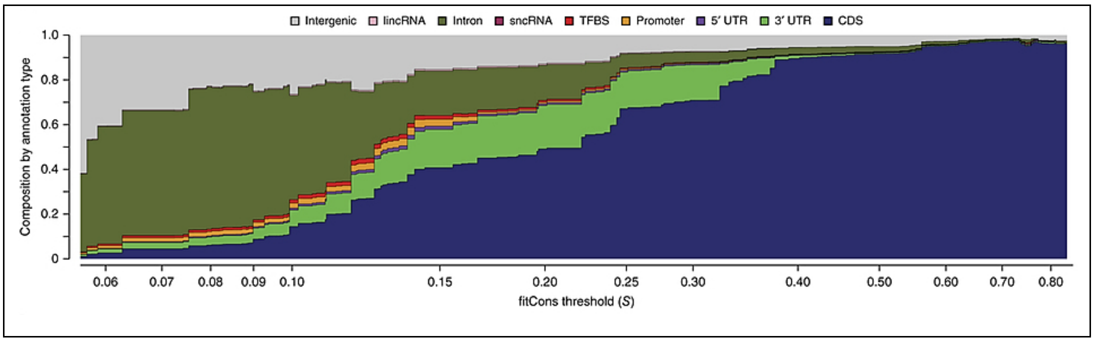


Figure 11 Composition of high-scoring genomic regions according to fitCons

Varying fitCons thresholds (S) are shown on the x-axis and the composition of various annotation types are shown on the y-axis[114]. Reprinted by permission from Macmillan Publishers Ltd: Nature Genetics 47:276-283, copyright 2015.

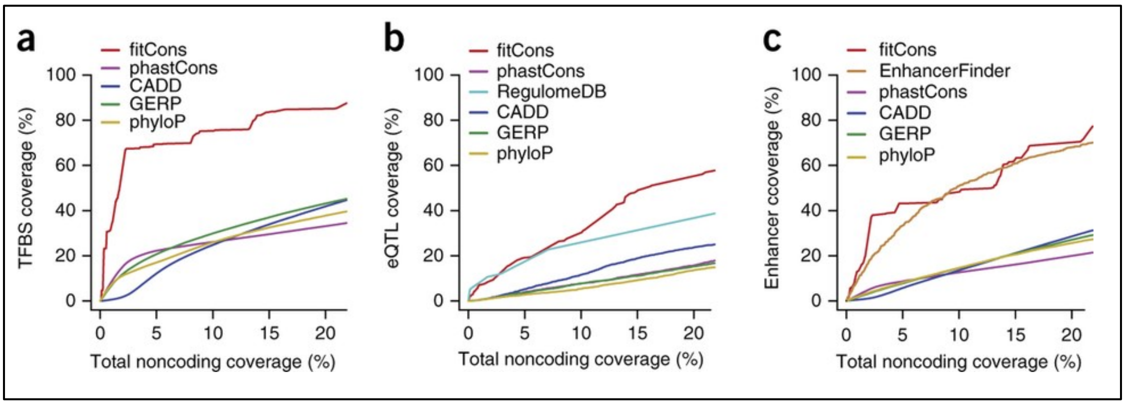


Figure 12 Coverage of active cis regulatory elements as a function of total coverage of the noncoding genome

Coverage of each type of element is shown as the score threshold is adjusted to alter the total coverage of noncoding sequences in the genome, excluding sites annotated as CDSs or UTRs[114]. Reprinted by permission from Macmillan Publishers Ltd: Nature Genetics 47:276-283, copyright 2015.

The ability of CADD and fitCons to accurately determine the functional importance of non-coding SNPs is therefore not comprehensive and manual assessment is therefore required. Unfortunately, neither CADD nor fitCons provide access to the original evidence used to produce the score which makes manual assessment challenging.

Additionally, algorithms, like ChromHMM[118], have been developed which classify regions into various chromatin states giving researchers the ability to quickly ascertain if a region is functionally relevant, however they do not combine other functionally important features such as transcription factor binding or DNase I HS and thus are still limited. Nevertheless ChromHMM classifications are a useful determinate of regulatory potential and can allow classification of up to 50 chromatin states by using imputation based on six epigenetic marks[119].

As technology improves and NGS costs reduce, more epigenomics data will undoubtedly be produced, in increasing numbers of cell types, including primary cells, and under different stimulatory conditions and the task of combining and, most importantly, interpreting this data will get increasingly more difficult. It is therefore important that researchers have access to these tools which allows them to make sense of the data and to inform subsequent experiments. However, there is still a need to be able to statistically aggregate and model related, complex data, including temporal and spatial data, to fully explore transcriptional regulation in the context of disease.

Despite the usefulness and innovative approach of ASSIMILATOR it has been superseded by other annotation tools, such as HaploReg, RegulomeDB and CADD. Due to the complexity of the publicly available epigenomic data that now exists, coupled with the knowledge of the importance of cell type, it is probably no longer valid to summarise this data across cell types. An enhanced approach would be to aggregate data by cell type and return the evidence for the most functionally important cell type using a scoring system such as RegulomeDB. Additionally, a method, such as ChromHMM as employed by fitCons, which could summarise at least elements (i.e. histone ChIP-Seq data) of the vast amount of data into well-defined functional classes before testing for functional importance would be beneficial from both a query time and resource utilisation perspective.

3 Vitamin D Response Element Enrichment in RA (Publication 3)

3.1 Background

Complex diseases are a combination of genetics and environmental factors and there is evidence that the environmental factors can increase the risk of disease through interactions with genes (GxE)[120]. This is because different genotypes can respond to environmental changes, such as physical shock (temperature) or chemical exposure, in different ways. Individuals carrying a ‘high-risk’ genotype do not necessarily develop disease but are more sensitive to an environmental factor which causes disease. For example individuals with fairer skin have a higher risk of developing skin cancer due to exposure to sunlight than darker skin individuals due to naturally lower levels of melanin[121].

However, most GWAS do not incorporate any environmental factors and rely on genetic evidence alone. This can be due to multiple reasons: firstly, environmental factors can be difficult to robustly define or measure; secondly, while GWAS have benefitted from increasingly lower costs and provide millions of results per individual, environmental factors can still be expensive and time consuming to collect and record thoroughly; thirdly, it is often unclear or unknown which environmental measure is required and therefore a specific hypothesis must be tested; and finally, statistical methods to detect GxE are less well defined and interaction analyses typically require four times the number of samples compared to analyses used to identify a main effect of similar magnitude[122].

Vitamin D is a steroid hormone involved in many biological processes including bone metabolism, muscle strength and modulation of the immune system[123]. The active form of vitamin D, 1,25-dihydroxyvitamin D3 (VitD$_3$), modulates its biological effects by binding to the nuclear vitamin D receptor (VDR) which can then act as a ligand-inducible transcription factor[124] and control more than 200 genes, including ones involved in regulation of cellular proliferation, differentiation, apoptosis and angiogenesis[123]. It achieves this by binding to specific elements in the genome, called vitamin D response elements (VDREs), and subsequently regulation of its target genes.

Vitamin D deficiency is common in RA[125] and as vitamin D is known to induce immunological tolerance[125], deficiency may disrupt this by inducing the development of disease. Vitamin D has also been shown to induce Tregs[126] and inhibit the production of proinflammatory cytokines[127], such as IL-2 and interferon-γ (IFN-γ), which in turn can cause a reduction in antigen presentation in antigen presenting cells (APCs)[128], thereby reducing T cell activation.

A study by Ramagopalan *et al.*[33] used chromatin immunoprecipitation followed by next-generation sequencing (ChIP-Seq) to characterise all VDREs in lymphoblastoid cell lines (LCLs) from two individuals of European ancestry before and after calcitriol (VitD$_3$) stimulation. The authors then tested for and showed significant enrichment of VDREs in known autoimmune disease (AID) loci, including RA ($p<0.001$). However, this study only tested 16 loci for RA and only 9 of these were confirmed to be associated with RA. Since its publication, the RA Immunochip study had also identified 48 non-HLA loci confirmed to be associated with RA susceptibility. Additionally, a study conducted in a T1D cohort identified two loci involved in vitamin

D metabolism (*DHCR7* and *CYP2R1*) which were associated with disease susceptibility.

3.2 Aims and achievements

The aims of this study were, firstly, to test for a potential enrichment of VDREs in RA loci by combining our genetic association data with the VDRE ChIP-Seq data to investigate whether the functional mechanism in RA associated regions acts through an interaction with vitamin D and secondly, to test variants previously associated with vitamin D levels and disease susceptibility in T1D and RA.

To achieve this, firstly a total of 2,776 VDREs identified after calcitriol stimulation were obtained from Ramagopalan *et al.*[33] and assigned to the nearest gene within 100kb. RA associations were assigned to genes either by GRAIL or by position and compared to the VDREs to determine the number of VDREs present in RA loci. To identify an enrichment of VDREs in RA loci, the average number of VDREs identified in 100,000 comparison sets of the same number of randomly selected loci was used to calculate the relative risk (RR) for RA loci. To test existing vitamin D loci associated with T1D, genotyping was carried out as part of the RA Immunochip study restricting the data to UK samples only.

Our study showed significant enrichment of VDREs in RA loci when associations were either assigned to genes by GRAIL (RR 5.50) or by position (RR 5.86) supporting a role for vitamin D in RA pathogenesis[32]. Additionally, evidence of association with the previously identified T1D locus, *DHCR7*, was also observed in RA (p=0.0008) providing further evidence supporting this conclusion.

3.3 Contribution to the literature

Many epidemiological studies have investigated the link between vitamin D deficiency and autoimmune diseases (AIDs), including T1D, multiple sclerosis (MS) and RA[129–134] but these studies have used either questionnaires or serum 25-hydroxyvitamin D (25(OH)D), the precursor to the active VitD$_3$ and have not incorporated genetic evidence. Likewise genetics studies have also been conducted to investigate vitamin D deficiency but have again focused on serum levels of 25(OH)D, candidate genes and RA risk[135,136]. While Ramagopalan *et al.*[33] decided to take a different approach and test for enrichment of VDREs, their approach for RA was limited by the genetic associations used and how VDREs were assigned to associated variants.

Our study improved on their analysis by utilising additional, validated loci and accounting for the number of VDREs attributable to each genetic association using a variety of methods. Using enrichment based methodology we have tested the potential environmental effect of vitamin D and RA susceptibility in an unbiased manner. This study, provides evidence that vitamin D deficiency is a cause of RA, although a recent study has suggested that lower levels of vitamin D present in severe RA patients is more likely a consequence rather than a cause and does not support the role of vitamin D supplementation as a direct therapeutic intervention[137].

Although at the time, our study was the most comprehensive to date, similarly to Ramagopalan *et al.*, it is now limited by the genetic associations used to test for enrichment as there are now 101 non-HLA RA associations[28]. A more comprehensive analysis would therefore be to repeat this method using all 101 RA associations currently known. This however would not be an ideal approach as methodologically, this study has also been superseded by other methods, such as Genomic Annotation Shifter (GoShifter)[70] and Mendelian randomisation[138].

As mentioned previously, GoShifter tests for enrichment of query features, such as SNPs, within a specific set of annotations, such as histone marks. GoShifter works by firstly, identifying all variants in LD ($r^2 > 0.8$) with each association to define a SNP region. Secondly, the observed overlap of each LD SNP with the annotations is determined. To produce a null distribution, each region is randomly shifted and the proportion of overlap is determined. This shifting process is repeated many times to generate a distribution. The significance of the overlap for the associated SNPs can then be determined by where in the distribution it lies relative to the random shifts. This process has an advantage over other methods as it maintains the genomic context for each region and the authors show that this yields more power compared to SNP matching methods (Figure 13). This method has been used to show that 88 RA SNPs from Okada *et al.* are enriched for H3K4me3 ChIP-Seq summits (±100bp) across all 118 cell types tested and specifically $CD4^+$ memory T-cells. This association remained significant after stratification on the other 117 cell types. This method could therefore be applied to test for enrichment of VDREs for SNPs associated with RA.

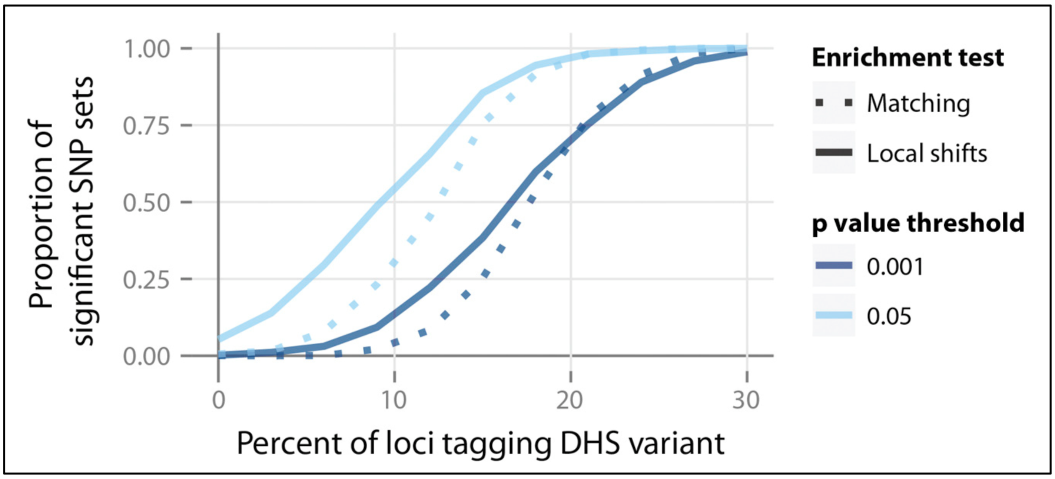


Figure 13 Comparison of power between GoShifter and the best-performing matching strategy

Two significance levels ($p < 0.05$ and $p < 0.01$) are shown[70]. Sets of 1,416 SNPs were generated with varying proportions within DNase I HS features (x-axis).

This approach, however, may not necessarily prove that vitamin D deficiency is a cause of RA as there are multiple additional factors affecting symptom onset, such as other risk variants, environmental factors and the interplay between them, which this method would not account for. To overcome these confounding factors, Mendelian randomisation has been proposed as a possible approach. First suggested in 1986 as an approach to show the relationship between cholesterol levels and cancer[139] and implemented by Gray and Wheatley in 1991 to study the effect of bone marrow transplants[138], Mendelian randomisation provides a similar study design as a randomised control trial (Figure 14) and uses genetic variants of known function to test the causal effect of an exposure on disease. For example, if you believe that low cholesterol (the exposure) causes cancer (the outcome), simply testing levels of cholesterol in cancer patients would not tell you if cholesterol was a cause of or due to cancer and a traditional randomised control trial would not be possible. SNPs associated with cholesterol levels can be tested for association in cases because alleles are inherited randomly, due to meiosis, and the presence of a particular allele or genotype in the population should be unrelated to any potential confounding factor. If low cholesterol causes cancer then the SNPs associated with low cholesterol level should be more common in cancer patients than controls; if not then it shows that cholesterol level is an effect of cancer.

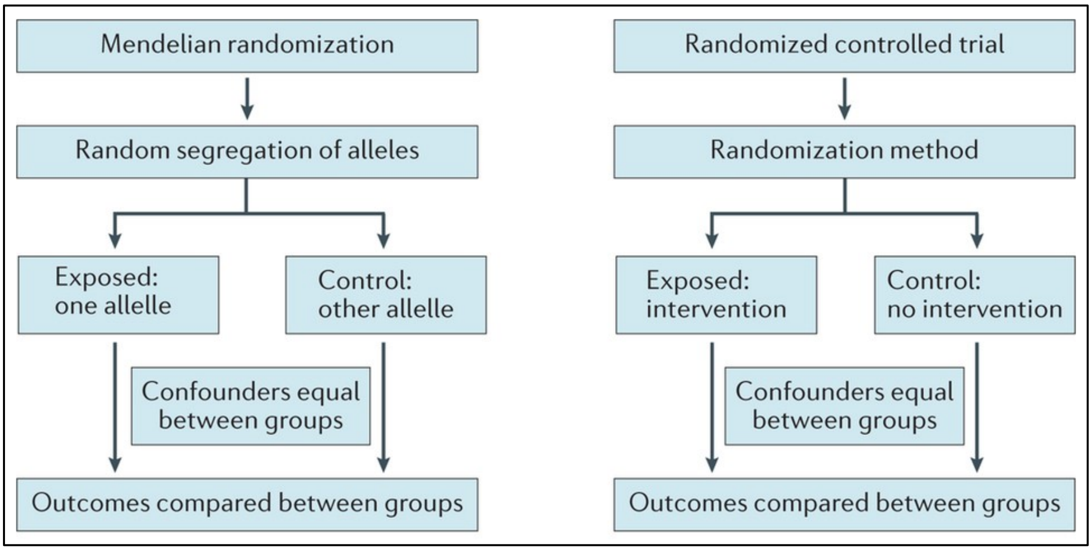

Mendelian randomization
Random segregation of alleles
Exposed: one allele
Control: other allele
Confounders equal between groups
Outcomes compared between groups
Randomized controlled trial
Randomization method
Exposed: intervention
Control: no intervention
Confounders equal between groups
Outcomes compared between groups

Figure 14 Comparison of the design of a Mendelian randomization study and a randomized controlled trial

In a randomised, controlled clinical trial, participants are randomly allocated into intervention (exposure) and control (no exposure) arms. In a Mendelian randomisation study, this is achieved through random segregation at meiosis. Both groups are equally exposed to confounders[140]. Reprinted by permission from Macmillan Publishers Ltd: Nature Reviews Rheumatology 12:486-496, copyright 2016.

The use of Mendelian randomisation studies has only increased recently due to the ever increasing amount of GWAS data and variants associated with multiple traits. For example, an adequately powered Mendelian randomisation study requires large datasets (>10,000 individuals)[141] which has only been possible for many traits with the increased application of GWAS. The method has been successfully applied to several relationships, such as coronary heart disease and C-reactive protein (CRP) levels, obesity and vitamin D levels and RA and levels of IL-1 and vitamin D. The findings of these studies are shown in Table 2.

Table 2 Example Mendelian randomisation studies

Exposure	Disease/outcome	Conclusion	Reference
C-reactive protein	Coronary heart disease	"C reactive protein concentration itself is unlikely to be even a modest causal factor in coronary heart disease"	C Reactive Protein Coronary Heart Disease Genetics Collaboration (CCGC)[142]
Vitamin D	Obesity	"that a higher BMI leads to lower 25(OH)D, while any effects of lower 25(OH)D increasing BMI are likely to be small"	Vimaleswaran *et al.*[143]
IL-1 inhibition	Cardiovascular risk /RA	"that long-term dual IL-1α/β inhibition could increase cardiovascular risk and, conversely, reduce the risk of development of rheumatoid arthritis"	Freitag *et al.*[144]
Vitamin D	RA	"the reported lower vitamin D levels in more severe RA are more likely to be consequence than cause"	Viatte *et al.*[137]

Mendelian randomisation therefore promises to be a powerful approach to link exposures, such as vitamin D, to causation in disease and indeed Viatte *et al.* has used this approach to show that lower vitamin D levels are more likely a consequence of RA[137]. However, these type of studies do have limitations. Firstly, any effect on the outcome must be as a direct result of the exposure and not due to any off target effects. Secondly, the robustness of the genetic association or associations must be ensured as these form the basis of subsequent tests. Since most of the genetic variants used in Mendelian randomisations will come from GWAS associations, the direct link of the association on the exposure must be ensured, otherwise more confounders will be introduced and any effect observed on the outcome may itself be due to an indirect link with the exposure. Additionally, it is usually not possible to include the entire genetic component of the exposure and gain sufficient power to identify a link. It is therefore important to ensure that the study design fully addresses these limitations and the power to achieve this will only increase as more genetic data is obtained and linked directly to particular traits.

4 Linking variants to target genes (Publications 4, 5 and 6)

4.1 Background

GWAS has now been successful in identifying over 100 genetic variants associated with susceptibility to RA. However many of these variants, like other complex disease associations, lie outside traditional protein coding regions suggesting that they have a role in gene regulation rather than directly affecting the protein produced. One of the most obvious ways which GWAS variants are likely to affect gene regulation, which is generally different in different cell types, is by altering the DNA binding motif of a transcriptional activator or repressor such as a transcription factor and indeed, there has been increasing evidence to suggest that gene expression and eQTLs can be cell type and stimulus specific[13,14].

Fairfax *et al.* studied the effect of stimuli on naïve CD14$^+$ monocytes by performing whole-genome genotyping and mRNA expression using microarrays. They used hierarchical clustering to investigate changes in monocyte expression after stimulation, with lipopolysaccharide (LPS), a component of gram negative bacteria which triggers Toll-like receptor (TLR) 4 signalling and IFN-γ, a cytokine which acts through the JAK-STAT pathway. Ye *et al.* primarily studied the effect of stimulation duration on CD4$^+$ T-cells using expression microarrays at 0, 0.75, 2, 4, 10, 24 and 48 hours, stimulating the TCR receptor alone, with anti-CD3/anti-CD28 beads, or in combination with conditions favouring T$_H$17 differentiation or IFN-β. Additionally, they followed up a subset of genes and conditions to show differences in inter-individual and population variability using a NanoString panel.

Specifically, they both found either eQTLs or expression of certain genes, which were shared across cell states, whereas for other genes, the eQTL or expression was only present after stimulation. These findings show the importance of performing eQTL or expression experiments in the correct cell state as results based on naïve cells may not represent a stimulated or disease relevant cell state or reflect the true state of cells used in related experiments.

Fairfax *et al.* also found eQTLs were effected by stimulus type, either showing only significance under one stimulatory condition or showing different effect sizes between stimulatory types, for example, *IL8* and *TRAF6* eQTLs were only significant in LPS and IFN-γ stimulated cells respectively. The same was also true for gene expression. Ye *et al.* found 289 and 270 genes which were only expressed upon co-stimulation with T$_H$17-biased conditions and IFN-β respectively (Figure 15a and b). This shows that while stimulation is an important consideration, which stimulation is just as important to ensure relevancy when comparing between experiments. Additionally, Ye *et al.* showed an inter-individual variation in expression between genes (Figure 15c), with some genes, such as *IL2* and *TNF* showing little variation whilst others, such as *IL3* and *IL17A* showed much more variation between individuals. These findings could be attributable to genotype or environment. Indeed the authors sampled a subset of individuals on different dates and showed that generally, T-cell response is reproducible, with common cis genetic effects accounting for ~25% and physiological effects for ~4% of the observed variation. Although an in-depth eQTL study would need to be performed to fully resolve the effect of genotype on gene expression in these samples.

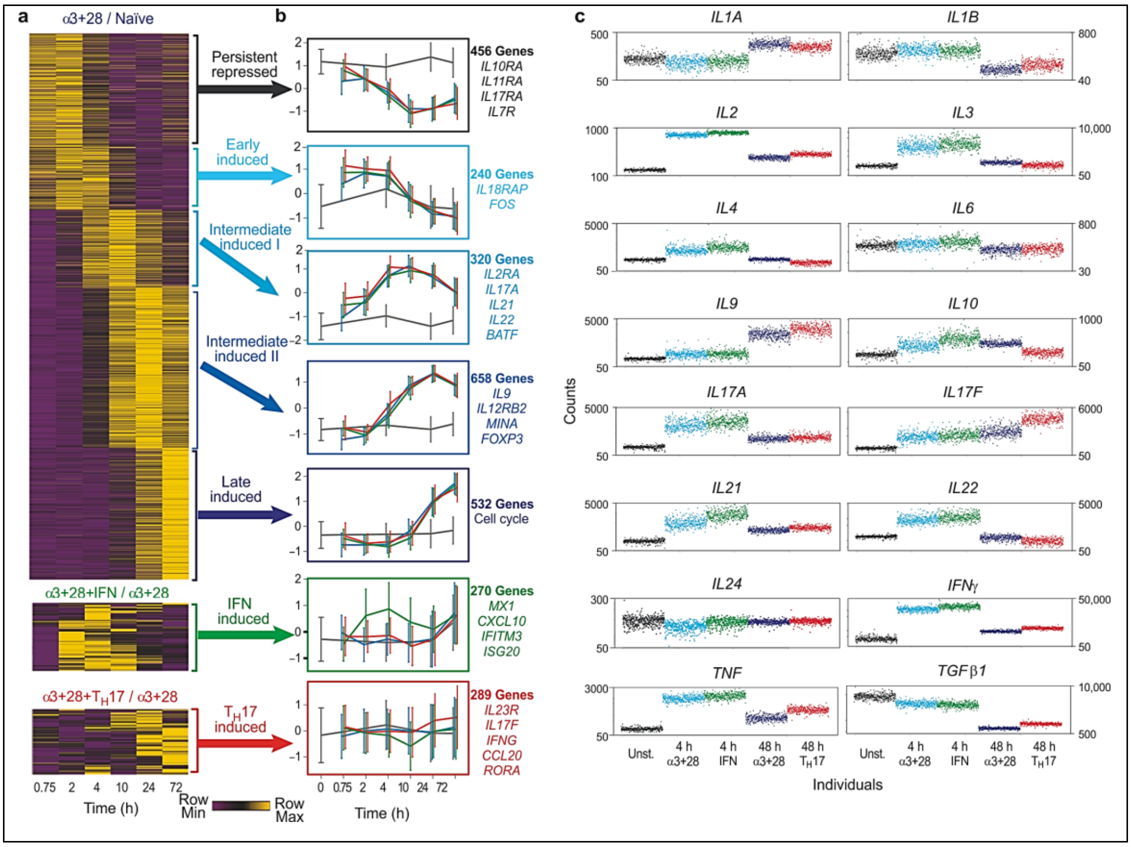


Figure 15 CD4⁺ T-cell time course gene expression profiles from Ye *et al.*

Cell states are colour coded as follows: black – naïve; blue – anti-CD3/anti-CD28 stimulated; red – TH17-biased co-stimulation and green – IFN-β co-stimulation. a) Clustering of genes across time points and stimulatory conditions. b) Expression profiles for each cluster identified. c) NanoString expression profiles of 16 cytokines showing cell state specificity, stimulatory duration specificity and inter-individual variation. From Ye *et al.*, Science 345:1254665 (2014). Reprinted with permission from AAAS.

Both studies found that changing the duration of stimulation affected not only, which genes were expressed (Figure 15b), but also how the expression of certain genes is affected by genotype such *LTA* and *TNF* (Figure 16) and while the time course performed by Ye *et al.* offered a more complete picture, Fairfax *et al.* were also able to identify the importance of duration of stimulation and its effect on gene expression, with additional influence by genotype. Importantly, they were also able to show that eQTLs specific to a cell state are found further away from the transcription start site (TSS) relative to those shared across cell states (Figure 17a), with stimulation specific eQTLs showing increased distance from the TSS (Figure 17b). Interestingly, for a minority of eQTLs, which were observed across conditions, the direction of effect was reversed between conditions (Figure 18). These findings highlight the complexity involved in gene regulation and suggest a potential mechanism of gene regulation which is dependant or defined by distance from the TSS.

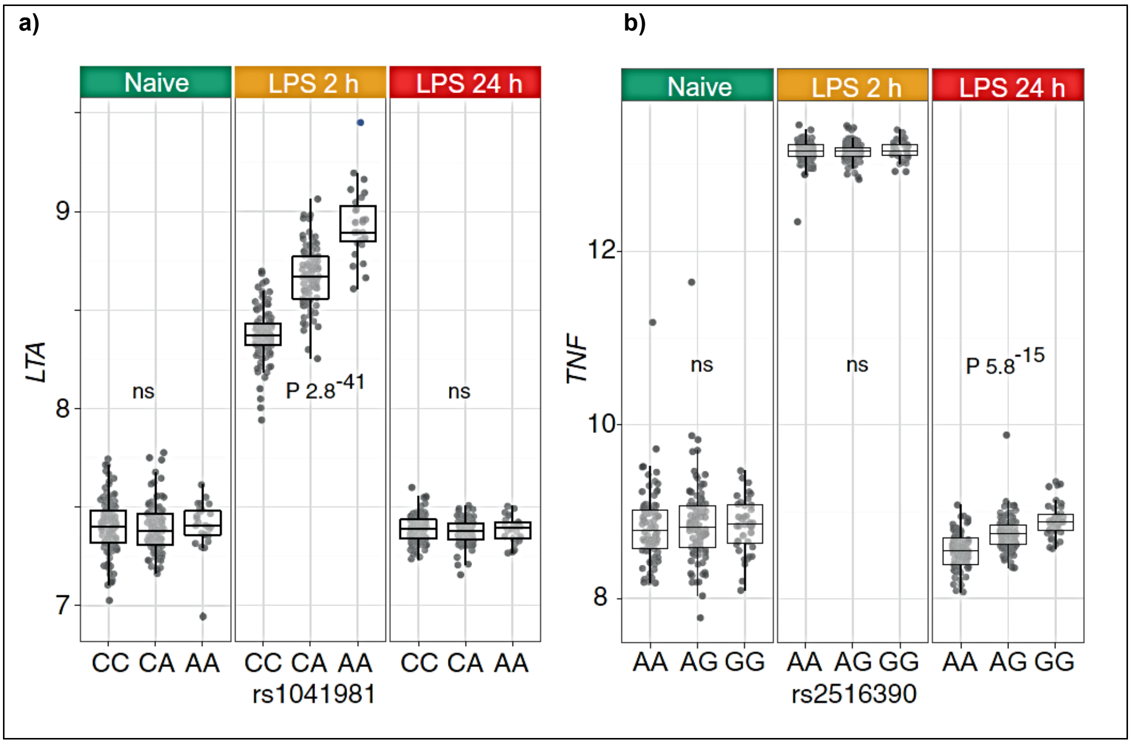


Figure 16 Duration of stimulation by LPS affects significance of certain eQTLs

Results are shown for a) *LTA* and b) *TNF*[13]. From Fairfax *et al.*, Science 343:1246949 (2014). Reprinted with permission from AAAS.

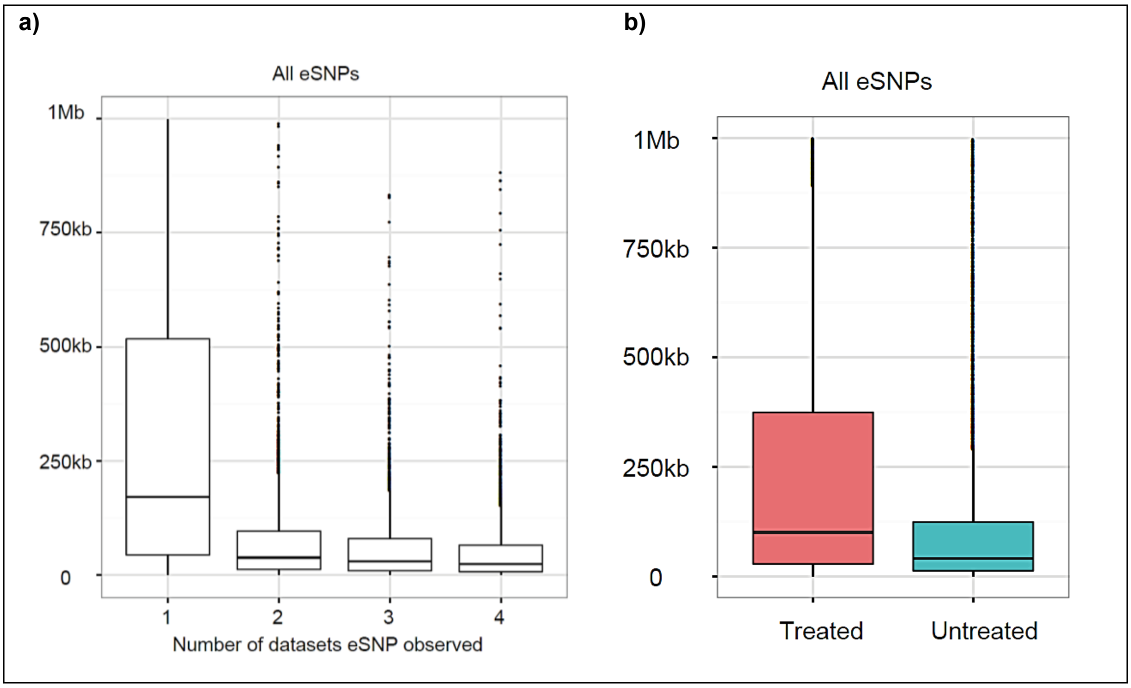


Figure 17 Distance of eQTL from TSS

Results are shown by a) number of cell states the eQTL is observed in and b) by cell treatment status[13]. From Fairfax *et al.*, Science 343:1246949 (2014). Reprinted with permission from AAAS.

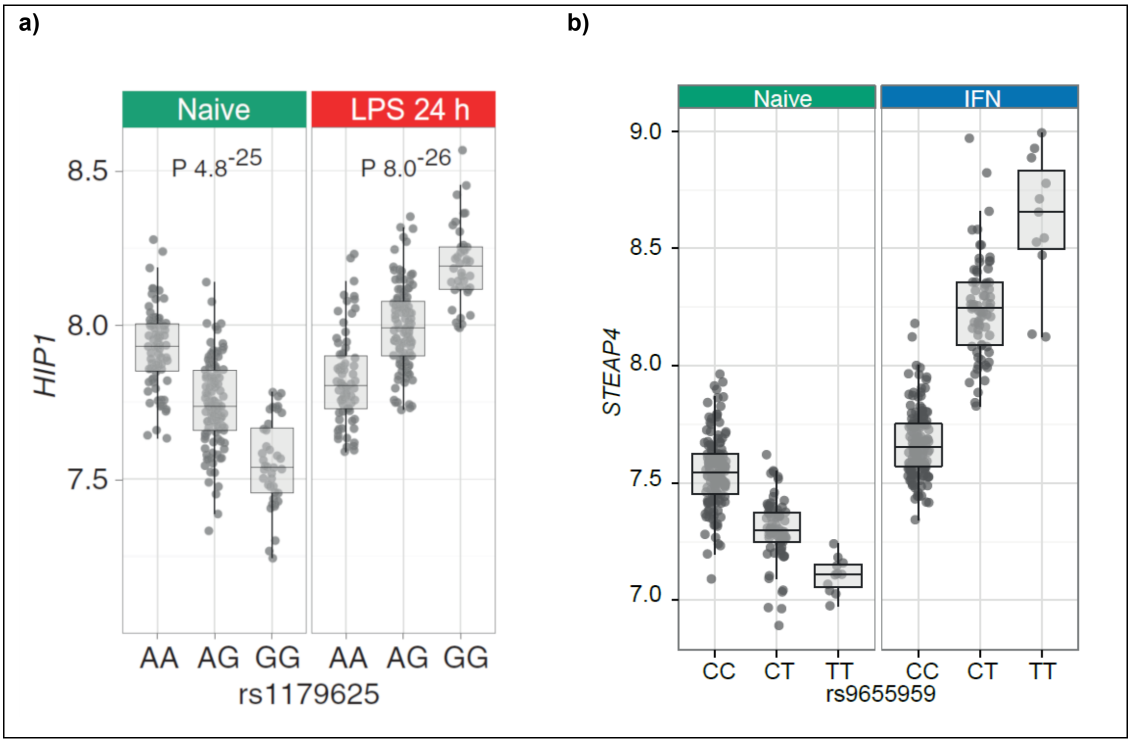


Figure 18 eQTLs showing opposing direction of effect

Results are shown for *HIP1* and *STEAP4* after stimulation with b) LPS and b) IFN-γ respectively[13]. From Fairfax *et al.*, Science 343:1246949 (2014). Reprinted with permission from AAAS.

GWAS variants have also been shown to be enriched for epigenetic marks indicative of enhancer elements which can also be cell type and stimulus specific. To fine map

GWAS associations and link them to transcription and cis-regulatory element annotations, Farh *et al.*[12] developed an algorithm, Probabilistic Identification of Causal SNPs (PICS), which uses the haplotype structure and pattern of associations at a locus, to estimate the probability of a SNP being causal. The PICS algorithm was applied to 21 autoimmune disease datasets and the authors showed that the majority (~90%) of candidate causal SNPs did not affect protein coding genes. Next they investigated the functions of these non-coding variants by mapping them to a set of specialised cis-regulatory elements, defined by H3 lysine 27 acetylation (H3K27ac), a mark indicative of active promoters and enhancers, for 56 individual cell types, including CD4$^+$ T-cells, Tregs, B-cells and monocytes. This revealed enrichment of the candidate enhancers in B-cells and T-cells (Figure 19). This finding highlights the differences in enhancers between cell types and also shows enrichment for stimulus dependant enhancers. Indeed, by coinciding PICS SNPs with cis-regulatory elements, Farh *et al.* predicted cell types contributing to disease (Figure 20).

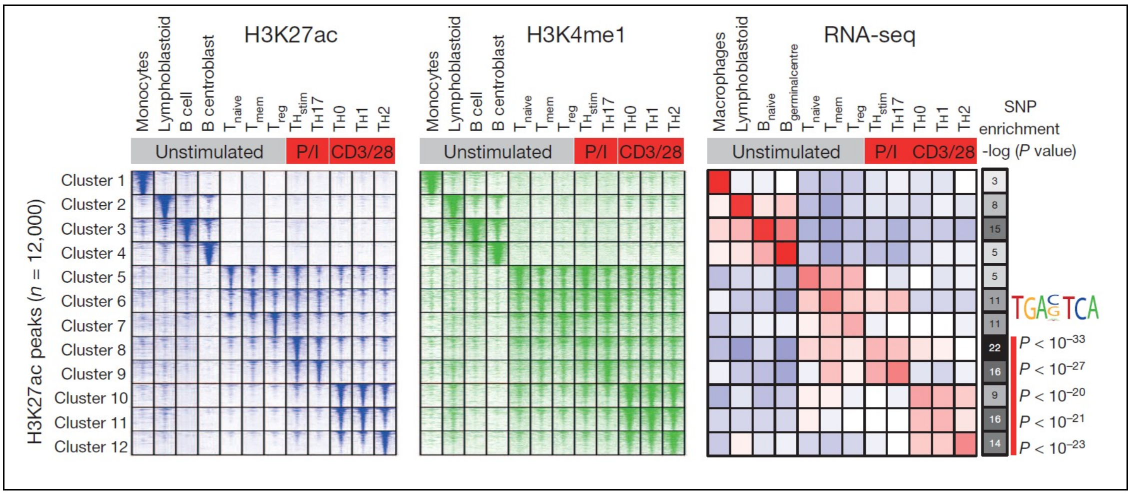


Figure 19 PICS enhancer mapping

Heatmaps for H3K27ac and H3K4me1 signals for 1,000 candidate enhancers (rows) in 12 immune cell types (columns). Enhancers are clustered by the cell type-specificity of their H3K27ac signals. The adjacent heatmap shows the average RNA-Seq expression for the genes nearest to the enhancers in each cluster. Greyscale (right) depicts the enrichment of PICS autoimmunity SNPs in each enhancer cluster[12]. Reprinted by permission from Macmillan Publishers Ltd: Nature 518:337-343, copyright 2015.

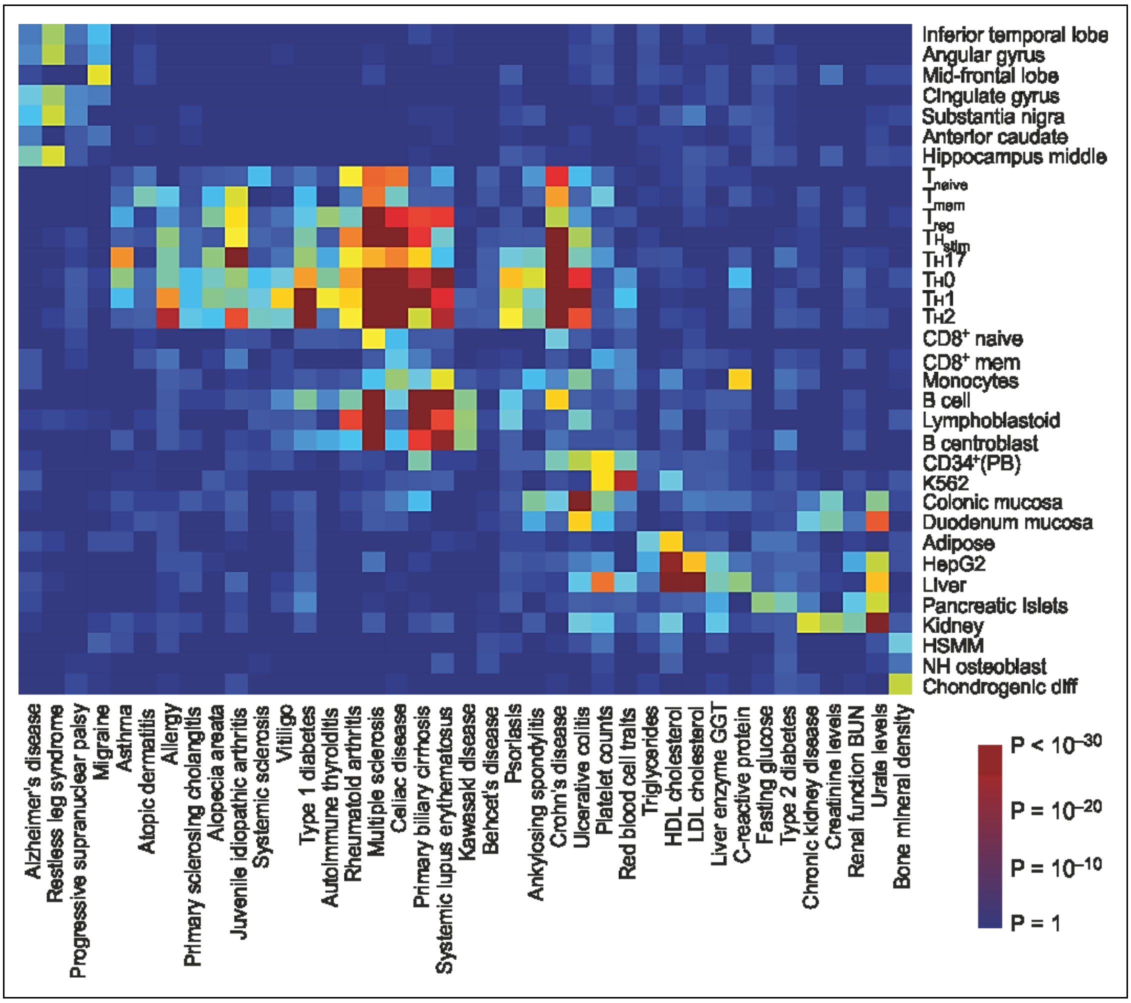


Figure 20 Heatmap showing cell type specificity of 39 human diseases in acetylated cis-regulatory elements of 33 cell types

Colour represents the P-value from 10-30 (dark red) to 1 (dark blue)[12]. Reprinted by permission from Macmillan Publishers Ltd: Nature 518:337-343, copyright 2015.

Additionally, they showed enrichment in super-enhancers, large regions with several enhancers in clusters, as well as evidence for different diseases mapping to distinct elements within a super-enhancer. For example, a candidate SNP lying in the *IL2RA* super-enhancer for MS has no effect on autoimmune thyroiditis risk, and vice versa for another candidate SNP, even though the SNPs are in close proximity. This suggests that some enhancer elements are specific to certain diseases and can effect, even a shared locus, in different ways. Interestingly, they also found GWAS SNPs associate with areas of the genome indicative of transcription factor occupancy, the specificity of which is dependent on disease and that many eQTL SNPs identified in peripheral blood do not correspond to enhancer elements (Figure 21), suggesting that many disease SNPs exhibit subtle and highly context-specific effects. These findings further highlight the cell type and stimulus type dependency of enhancer elements effecting gene regulation in disease.

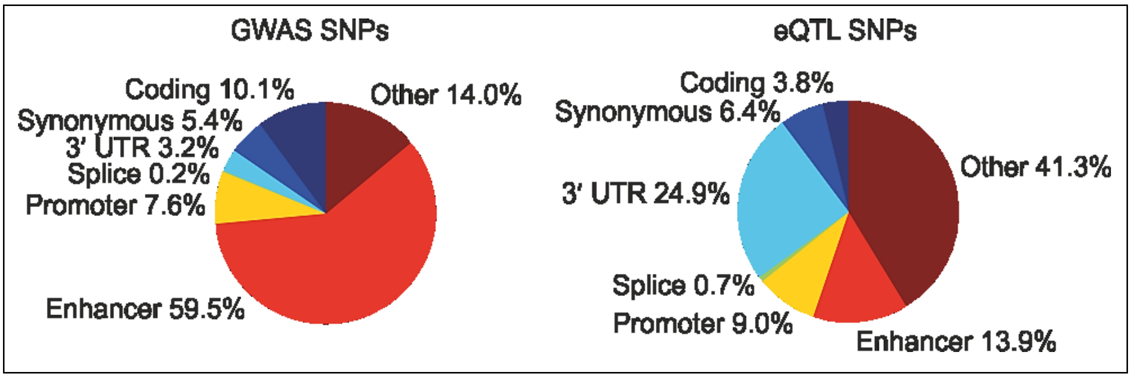


Figure 21 Functional effects of disease variants on gene expression

Pie charts showing of proportion of PICS SNPs (left) and eQTLs (right) explained by the genomic features shown[12]. Reprinted by permission from Macmillan Publishers Ltd: Nature 518:337-343, copyright 2015.

Since a large proportion of GWAS SNPs are found outside protein coding genes, it is imperative to identify which gene or genes they effect. Variants identified by GWAS have traditionally been annotated to the closest most biologically relevant genes and while this strategy may seem sensible, it could also be incorrectly implicating genes which are not involved in disease or masking additional effects with other genes. For example, the locus containing the *CTLA4*, *CD28* and *ICOS* genes contains two SNPs independently associated with RA, one of which has been assigned to *CD28*, the other to *CTLA4*[28] based on biological plausibility (Figure 22). However, all three genes are involved in T-cell activation and therefore represent ideal candidates for RA. There is therefore the possibility that either all three genes are regulated by these SNPs, and operate together to affect T-cell activation, only one is regulated or indeed, the two candidates assigned are the functionally relevant genes for RA.

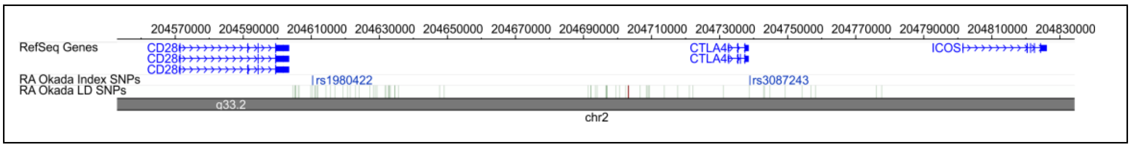


Figure 22 CD28-CTLA4-ICOS locus

Lead SNP associations are shown as well as SNPs in LD ($r^2 \geq 0.8$).

Furthermore, Musunuru *et al.* showed a SNP, rs12740374, associated with levels of low-density lipoprotein cholesterol (LDL-C), which is located within the 3' UTR of the *CELSR2* gene, actual regulates the expression of the *SORT1* gene[145]. Additionally, Davison *et al.* showed that SNPs, located predominately in intron 19 of the *CLEC16A* gene, and associated with T1D and MS, modify the expression of the *DEXI* gene using chromosome conformation capture (3C)[146].

An increasing number of studies, including Davison *et al.* have also shown that enhancers do not necessarily regulate the nearest gene (Figure 23a). In a study investigating the pilot ENCODE regions, Sanyal *et al.* showed that only 7% of elements regulate the nearest gene[147]. Additionally, elements were shown to regulate genes located some distance away, with a peak distance of 120kb, although further distances, up to 1.5Mb or more, have also been observed[148–150]. This long-range gene regulation, is achieved through chromatin looping (Figure 23b), thought to be mediated by cohesin and other protein complexes[151], which brings distant genomic regions into close proximity to regulate expression in a cell type and stimulus specific manner (Figure 23c).

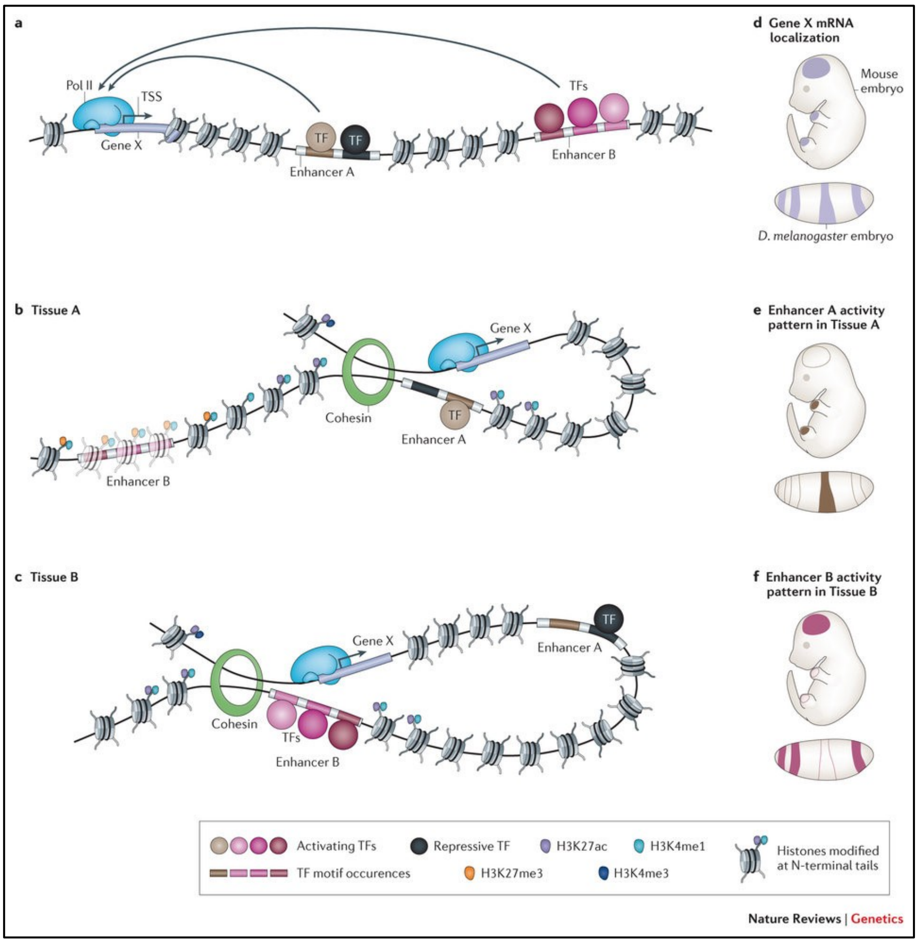


Figure 23 Long-range gene regulation

(a) Enhancers are distinct regions which bind transcription factors (TFs). These can be located at any distance from their target genes. However, when active (b), enhancers can be brought close to and interact with their target gene allowing them to regulate expression. These interactions can be tissue specific (c). (d-f) Patterns of gene expression. Source: Shlyueva *et al.*[101]. Reprinted by permission from Macmillan Publishers Ltd: Nature Reviews Genetics 15:272-286, copyright 2014.

Chromosome conformation capture methods, such as 3C, 4C, 5C and Hi-C, identify regions of the genome, such as enhancers and promoters, which physically interact in the nucleus. This is achieved using formaldehyde to crosslink the two interacting regions together, preserving any physical interactions present in the nucleus, followed by digestion with a restriction enzyme. The 'free ends' produced by the digest are then ligated together such that the interacting regions form a ligation junction which can be detected by assays designed for the particular technique. These early chromosome conformation capture techniques, such as 3C, 4C and 5C were successful in identifying chromatin interactions. Using 3C, Tolhuis *et al.* investigated interactions between the β-globin locus and the locus control region (LCR) in mouse[152]. They showed that in erythroid cells, the LCR, located 40-60kb away, comes into close contact with active β-globin genes to control their expression. Furthermore, Stadhouders *et al.* showed that an intergenic region interacts with the promoter of the *Myb* gene to up-regulate gene expression in proliferating cells[153]. However, these methods were low throughput and interaction targets had to be considered *a priori*.

Later, chromosome conformation capture was coupled with next-generation sequencing (Hi-C) to provide researchers with a high-throughput, hypothesis-free way to investigate chromatin interactions. Hi-C has been used to study the three-dimensional structure of the genome, identifying large, megabase-sized local chromatin interaction domains, such as topologically associated domains (TADs), large contiguous regions of the genome which associate in the nucleus and partition the genome into discrete domains[154]. TAD boundaries have been shown to be enriched for the insulator binding protein CTCF, housekeeping genes, transfer RNAs and short interspersed element (SINE) retrotransposons. This suggests they may have a role in establishing the three-dimensional structure of the genome. They are conserved across cell types and are highly species specific. The frequency of chromatin interactions within the TAD boundaries is much higher compared to across TAD boundaries. This may have implications on how GWAS SNPs regulate genes, predominantly affecting genes within the same TAD allowing discrete control of clusters of functionally related genes. Additionally, disruptions of TADs in the human WNT6/IHH/EPHA4/PAX3 locus have been linked to various limb malformations, such as brachydactyly and polydactyly[155], showing the functional importance of TADs and how variations within TAD boundaries could lead to a disease phenotype. However, Hi-C suffers from limited resolution and cannot resolve fine promoter-enhancer interactions.

Capture Hi-C was recently developed[156] to provide researchers with a high-throughput method to study interactions at high resolution for a defined set of targets, such as promoter or enhancer regions. The first step in a Capture Hi-C experiment is to generate a standard Hi-C library containing all the interactions present in the original sample (Figure 24). Using sequence capture technology[157], RNA baits, which are specifically designed to the targeted restriction fragment ends, are used to capture all the interactions involving the desired restriction fragments. These are then purified, enriched and sequenced at a much higher depth relative to a Hi-C library and can result in between a 19 to 130 fold enrichment over a Hi-C library depending on the number of targets selected and sequencing depth[156,158,159].

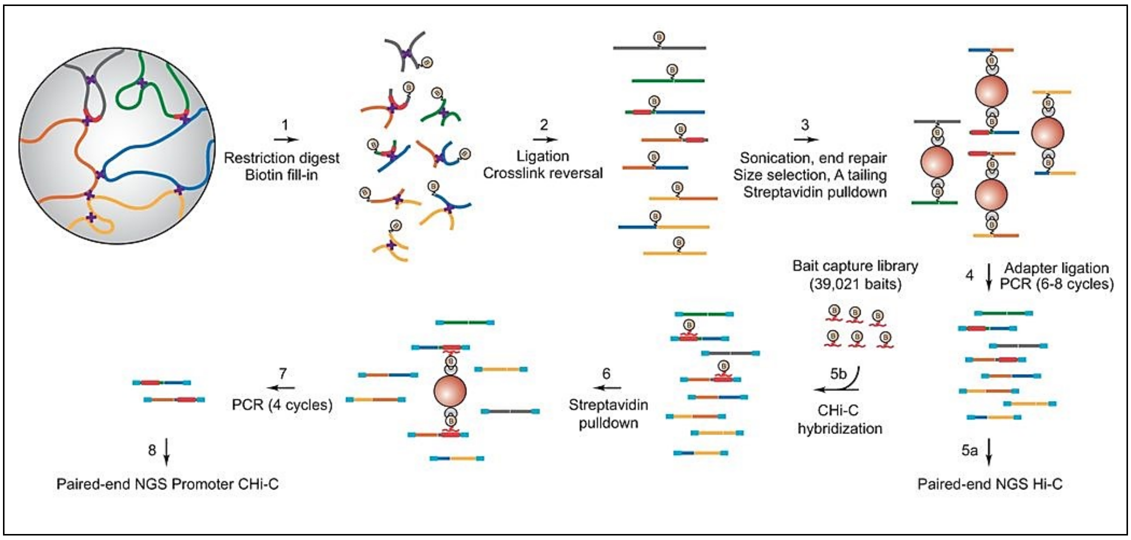


Figure 24 Capture Hi-C Overview

The first step in Capture Hi-C is to generate a standard Hi-C library by cross-linking DNA, digestion, re-ligation, reversal of cross-links, followed by purification. This provides a library containing all interactions in the cell. After adaptor ligation, RNA baits, designed to restriction fragment ends, are used to capture interactions specific to the targeted regions. This is followed by PCR and paired-end sequencing. Source: Schoenfelder *et al.*[156]

This method was initially used to study the chromatin interactions involving regions associated with breast cancer[159] which demonstrated the power of Capture Hi-C to identify high-resolution interaction maps for three breast cancer gene deserts mapping to 2q35, 8q24.21 and 9p31.2. The authors used a low LD cut-off (r^2>0.1) with the associated SNPs to define the regions to target, with the aim to identify interactions between regulatory elements and protein coding genes which could potentially be hundreds of kilobases apart. They identified 27 and 45 significant interactions for the BT483 and SUM44 breast cancer cell lines respectively. The majority of the interactions identified were tissue specific but there was also evidence of some interactions being common across the cell types studied and interactions were identified between both protein coding genes and long non-coding RNAs (lncRNAs). Additionally, using Hi-C data for one of the cell lines, they were able to show a 30-60 fold enrichment of the target loci by incorporating the sequence capture step.

Capture Hi-C was also used to study 14 loci associated with colorectal cancer[158] and also identified complex interaction networks and multiple long range interactions. Similar to Dryden *et al.* the authors defined capture regions using a low LD cut-off (r^2>0.2) and obtained an enrichment in excess of 130 fold over Hi-C data. Their data not only confirmed documented interactions, such as the one between rs6983267, a colorectal cancer risk loci, and the *MYC* gene, but also novel interactions involving plausible biological candidates, such as *CCAT1* and *CCAT2* which, together with *MYC*, suggest a network involving Wnt-feedback signalling.

These studies focused on a small number of large regions to test specific hypotheses related to cancer risk but do not offer a genome-wide method to provide a systematic and unbiased approach to study the chromatin interactions influencing gene regulation for thousands of targets, such as promoters. This led Mifsud *et al.* to design baits to study whole-genome promoter-enhancer interactions in two human cell lines by targeting all restriction fragments overlapping the promoters of Ensembl transcripts[160]. The authors discovered that the majority of interactions were between promoter and 'non-promoter' fragments, promoters would typically interact with tens of other ends, irrespective of transcriptional activity and other ends would interact with one or two promoter fragments. These results suggest that gene regulation involves a complex interplay between multiple genomic regions and that regulatory elements are shared between genes. Additionally, they also discovered that GWAS SNPs are enriched in fragments which interact with promoters, strengthening the role of GWAS SNPs in gene regulation.

In addition, promoter-enhancer interactions in mouse embryonic stem cells (ESCs) and mouse foetal liver cells (FLCs) have been investigated using this method[156] and the authors found similar observations as in humans. Their data also showed that genes with higher numbers of enhancer interactions tended to be enriched in developmental pathways for ESCs and metabolic pathways for FLCs supporting the hypothesis that the three-dimensional promoter-enhancer landscape is highly cell type specific. While whole-genome promoter capture experiments, such as these, offer a comprehensive view of genome regulation at high-resolution, a disease focused approach would yield interactions which are specific to disease associated loci, helping to inform gene candidature.

Similar methods to Capture Hi-C have been developed that incorporate a sequence capture step to 3C libraries, Targeted Chromatin Capture (T2C)[161] and Capture-C[162]. T2C follows essentially the same method as Capture Hi-C but is based on a 3C library and as such does not employ the enrichment step for biotinylated ligation products or di-tags. This step is crucial in removing 'invalid' interactions and DNA fragments left after sonication that would otherwise be sequenced unnecessarily. Capture-C, like T2C, suffers from the lack of enrichment for valid di-tags but does allow better resolution than both Capture Hi-C and T2C as it utilises a 4bp cutter as opposed to a 6bp cutter. Whilst no studies have performed a direct comparison of these methods, it is thought that the addition of the enrichment step in Capture Hi-C would offer superior signal to noise ratio and surpass the advantage of added resolution[158]. Additionally, other studies have employed 4bp cutters to Hi-C libraries[163] and, although this has yet to be applied to Capture Hi-C libraries, it therefore has the potential to offer both the benefit of high-resolution interactions and low signal to noise ratio. This approach does however also cause added statistical analysis considerations.

While Capture Hi-C does not offer a truly hypothesis free way to study chromatin interactions, it provides researchers with the ability to identify all chromatin interactions involving thousands of target fragments in a cost-effective, high-throughput manner. Additionally, the flexibility of this approach provides opportunities to study a wide variety of potential targets and aims.

4.2 Aims and achievements

The aim of this work was to use Capture Hi-C to characterise the physical interactions of associated loci for four autoimmune diseases: RA, JIA, PsA and T1D, with the objective of linking associated variants with causal genes. Uniquely, this was achieved using two Capture Hi-C experiments: the first targeted the associated regions for each disease, defined by LD; the second targeted promoters of genes within 500kb of each lead association (Index SNP) (Figure 25).

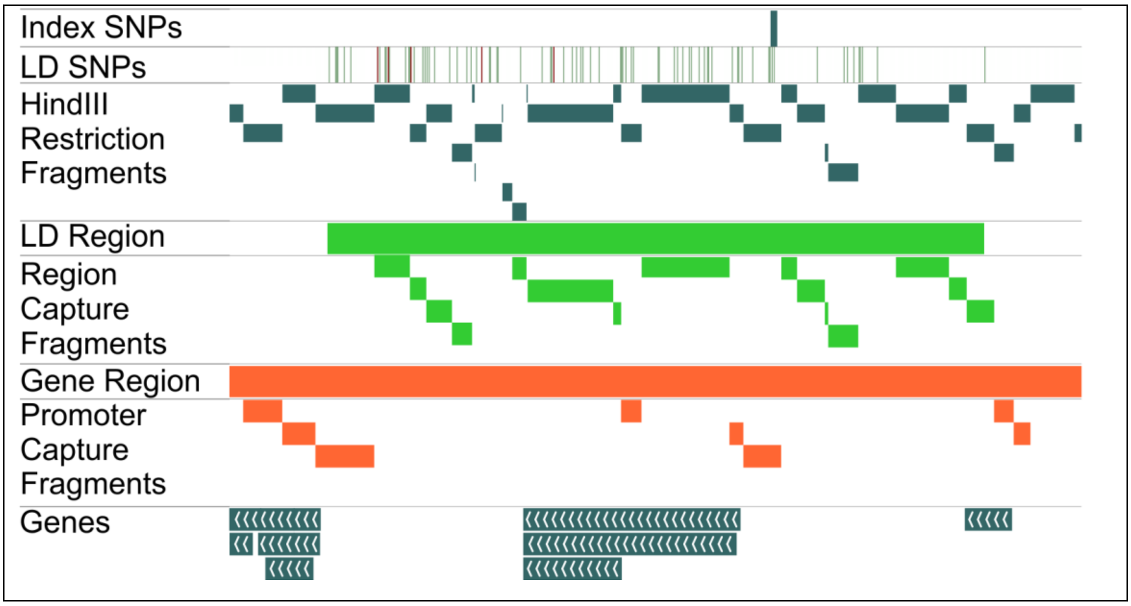


Figure 25 Capture Hi-C Experimental Design

HindIII restriction fragments targeted in the region capture are shown in green. HindIII restriction fragments targeted in the promoter capture experiment are shown in orange.

This study provides compelling evidence that GWAS SNPs may regulate genes located some distance away, SNPs associated with different AIDs may well regulate the same genes with different enhancer mechanisms and a number of interactions also show evidence of cell type specificity[34].

Following on from this study, comprehensive analysis of one RA locus, 6q23, revealed a complex pattern of interactions, implicating multiple immune related genes, such as *TNFAIP3*, *IL20RA*, *IFNGR1* and *IL22RA2*. Additional work on this region has confirmed these interactions, obtained bioinformatics evidence to narrow down the potential causal SNPs and shown allele specific histone marks and binding of the NFκB transcription factor. This work is the first study to comprehensively interrogate the chromatin interactions within this region and has highlighted the importance of gene assignment for translating GWAS findings to improve our knowledge of disease mechanisms and identify potential therapeutic targets.

This work on the 6q23 region also led to the identification of chromatin interactions involving regions not associated with RA and not in LD with these associations. Instead, these regions contained variants uniquely associated with multiple sclerosis (MS), an autoimmune disease affecting the central nervous system. It was therefore reasoned that our Capture Hi-C data could be used to investigate the mechanisms specifically affecting MS at this locus. The aim of this study was therefore to link the MS associations to potentially causal genes using Capture Hi-C data within this region and refine variants further using bioinformatics. While this study identified the reported GWAS genes as potential candidates, it also identified other related genes suggesting that MS associated variants could regulate not just one but multiple causal genes. Indeed, this work identified two clusters of chromatin interactions involving four lead MS associations within this region: one containing neurologically related genes and the other containing immunologically related genes. There is also evidence that independent disease associations interact with each other suggesting a complex regulatory mechanism where multiple regions associated with MS act cooperatively to regulate the expression of several genes. These findings could help us to understand the mechanisms of disease and also suggest potential novel therapeutic targets.

4.3 Contribution to the literature

This application of the Capture Hi-C method is the first to target the full known genetic component of four related AIDs at a much higher depth of sequencing (average 10,000 interactions per restriction fragment) compared to previous studies. Additionally, our unique, complementary study design allowed us to investigate chromatin interactions in a comprehensive, self-validating manner. This complementary approach is now being utilised by other studies to validate interactions observed in whole-genome promoter capture experiments[164] and offers a robust high-throughput method to confirm findings.

We have redefined how GWAS variations are assigned to genes, showing that it is often more complicated than simply the closest gene and that Capture Hi-C can be used to interrogate a large number of GWAS loci in a systematic and unbiased manner to identify potential gene targets and to further the understanding of complex diseases.

Our data has identified both existing and novel potential gene targets of disease associations giving the potential to inform future experiments to undercover the molecular mechanisms underpinning disease. Indeed, for the 6q23 region, this has already transpired as interactions identified in these experiments have already been utilised and expanded on to increase our understanding of how the genetic associations in the region not only interact with their target genes but also each other and between related diseases. Our work in the 6q23 region has also provided support for a new anti-IL-20 therapy which has been shown to be effective in the treatment of RA and psoriasis[165,166], showing that this method could be effective in identifying other novel or existing drug targets. Further work on our Capture Hi-C data has identified

41 genes which are targets for 109 existing drugs for RA alone, of these only nine are currently used in the treatment of RA[96].

While our work has focused on the four AIDs previously mentioned, it has the potential to also inform gene regulation in other related AIDs as considerable overlap has been observed between AIDs. For example, the 6q23 region also contains multiple associations to other AIDs such as MS, celiac and SLE and, while not targeted directly, our data may provide insight into genetic associations with these diseases. Further exploration of this region has shown that variants which are only associated with MS, interact with two regions, the first implicating neurologically related genes including *AHI1*, *SGK1* and *BCLAF1* and the second implicating immunologically related genes such as *IL20RA*, *IL22RA2*, *IFNGR1* and *TNFAIP3*[36].

Despite the success of this and similar studies, it is clear that the interactions identified are highly cell type and even stimulus specific and further work will be required to fully explore GWAS associations to identify the genes they regulate. Future Capture Hi-C studies should be conducted in primary cells, preferably patients, to fully explore the regulatory mechanisms which exist and underpin disease. This approach, coupled with matching eQTL data, would help to resolve which genes are affected by GWAS associations and how they act to cause disease.

5 Discussion

It is clear that whilst GWAS has been tremendously successful in identifying variants associated with common complex diseases, not least in RA, it is only the beginning in understanding disease and how these common variants act together with each other and the environment to increase an individual's risk to develop disease. Improvements in genomics technologies, increased sample sizes and larger, more collaborative research has driven this success but has also driven the evolution of bioinformatics. Only up to relatively recently, the initial GWAS generally did not include any bioinformatic analysis and simply presented the significant associations annotated to the closest or most biologically plausible gene. As GWAS became more and more common, bioinformatic analysis was gradually introduced and now comprises a substantial portion of GWAS publications. For example, only approximately one third of the publication by Okada *et al.* discussed the association results and trans-ethnic analysis. The remainder of the article described further bioinformatics analyses including epigenetic marks, functional annotation, overlap with disease, pathway and PPI networks and drug target evaluation.

Once it was clear that many of the genetic associations were outside of protein coding regions, the challenge was to determine how these variants could influence disease susceptibility. The obvious mechanism was gene regulation and projects such as ENCODE and the epigenomics roadmap project allowed researchers to annotate potential functional variants with regulatory elements, such as histone marks and DNase I HS sites. This presented a new challenge for researchers due to the wealth of data available and limited methods to query, integrate and investigate the functional elements in relation to disease associations. This led to the development of several bioinformatics tools to integrate this data to allow researchers to fully utilise the resource.

Furthermore, many techniques to study the regulation of genes, either by measuring the expression directly or identifying regulatory elements, require sequencing to produce high-throughput and cost efficient data. Prior to the development and widespread adoption of next generation sequencing (NGS), Sanger sequencing was used. This technology was low-throughput and expensive, costing around $5,000 per megabase (Mb) in 2001 (the completion of the human genome sequence[167,168], ~$95 million/genome) and although dropping significantly to around $400/Mb in 2007, still suffered from poor throughput and technological limitations (Figure 26). Importantly, sequencing costs within this period roughly followed Moore's law, an observation that compute power roughly doubles every year. However, with the advent of NGS in 2008, it triggered not only a rapid decline in the cost of sequencing, falling from $0.52/Mb in 2010 to $0.014/Mb or $1,245 per genome in 2015 and surpassing Moore's law, but also a rapid increase in the rate of sequence production. As the cost continued to drop and throughput continued to increase, NGS was applied to more techniques, such as RNA-Seq, ChIP-Seq and Hi-C, allowing researchers to study genome-wide regulation for the first time. This rapid increase of data coupled with the range of techniques, has cemented the role of bioinformatics in research and also has led to the establishment of many companies offering specific bioinformatics analysis solutions. Indeed, bioinformatics is now commonplace and many researchers are required to have at least a basic understanding of bioinformatics knowledge.

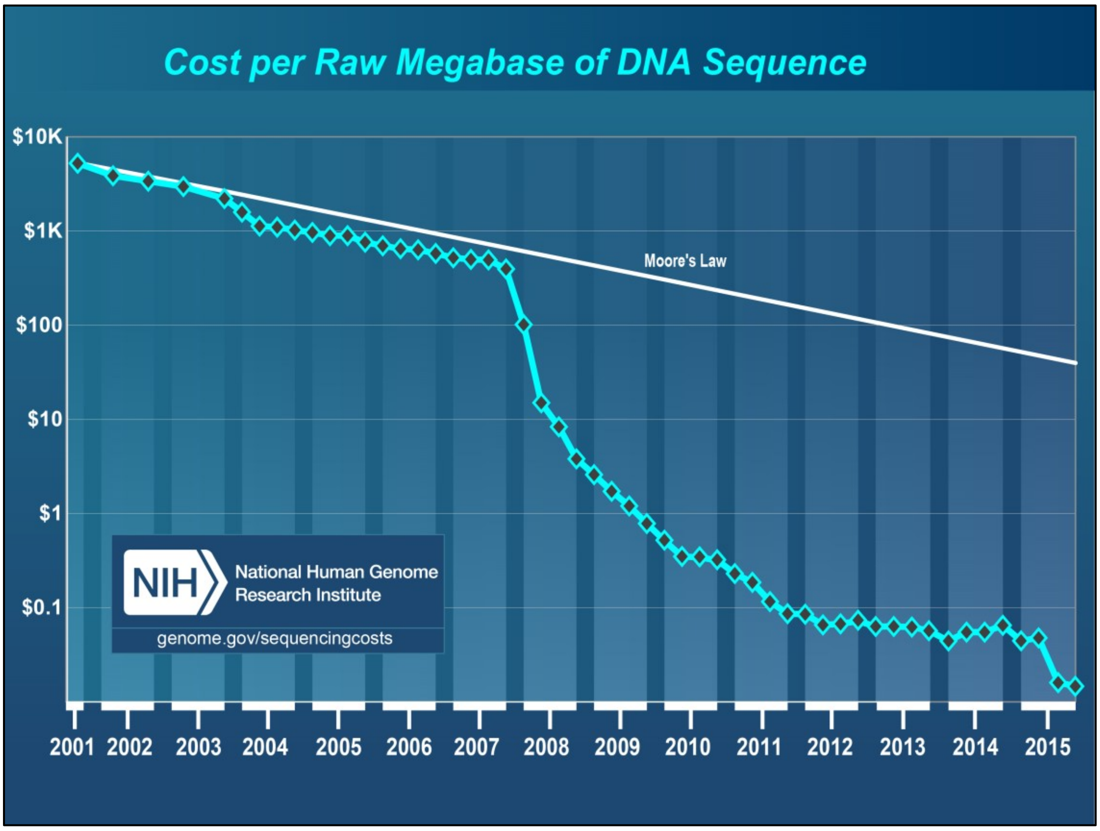


Figure 26 Cost per raw megabase of sequencing in US dollars between 2001 and 2015

Prices between 2001 and 2007 are for Sanger sequencing. From 2008 onwards, costs are based on NGS. Source: https://www.genome.gov/sequencingcostsdata/.

Importantly, since the introduction of NGS, cost and rate of sequence data generation has surpassed developments in compute power. This has meant that data analysis has moved away from large workstations and small clusters to large shared compute farms with thousands of cores and large amounts of memory. In addition, this has meant that analysis solutions require complex multi-core solutions, efficient algorithms and efficient sharing of data between processes. Almost all NGS tools employ some sort of multi-threading allowing them to utilise a set number of cores to reduce the overall compute time needed.

The increase in the use of newer techniques, such as single cell RNA-Seq and mass cytometry by time of flight (CyTOF) has provided new challenges to the field of bioinformatics. Single cell RNA-Seq has shown that a homogeneous population of cells can actually represent multiple cell sub-types, each exhibiting slightly different expression profiles and reacting differently to their environment. Macosko *et al.* showed that mouse retinal cells are comprised of 39 transcriptionally different sub-types using novel combinations of clustering techniques[169]. Following a similar trend as other NGS techniques, such as RNA-Seq, Pollen *et al.* showed that sub-types of developing cortex cells could be identified with as few as 10,000 reads per cell, compared to the 100,000 minimum read limit imposed by Macosko *et al.*, although to obtain finer distinctions between categories, 50,000 reads per cell were necessary[170]. This does however highlight how techniques can develop and the importance of bioinformatics to achieve this.

CyTOF provides a different challenge to single cell RNA-Seq as no sequence data is generated. Instead it uses mass spectrometry to measure more than forty cell surface markers on individual cells. Cell surface markers are labelled with a unique combination of metal-tagged reagents, such as antibodies. The time taken from excitation to detection allows the mass to be determined and therefore the corresponding antibody and the signal strength represents the frequency of markers on the cell. This technique has the ability to profile individual cells of a heterogeneous population and classify them into sub-types. However, analysis of the data is complex and requires a host of bioinformatics solutions, including new clustering methods, cloud based analysis solutions and visualisation methods.

Perhaps as significant a task as developing novel methods to evaluate the vast amounts of data being currently generated is the ability to combine these ‘omics data sets and to compare across cell types. The bioinformatics, and analysis fields, are now moving into areas of how to handle the enormous data sets, how to integrate the multi-level data and how to determine sensible conclusions from this data. For example, determining how variation relates to transcription factor binding, gene expression and protein levels, has the capacity to infer cause and effect, and the mechanism by which an associated variant increases risk of disease. This is obviously not trivial, requiring novel analysis tools, powerful computing resources and robust statistics.

The work presented here has used high level bioinformatics to solve specific research problems. Firstly to assist researchers to integrate a vast array of data to functionally annotate and select potential causal SNPs from GWAS associations. Secondly, to test for enrichment of specific genomic features, VDREs, in SNPs associated with RA to ascertain the involvement of vitamin D in RA pathogenesis. Thirdly, to analyse the complex three-dimensional chromatin interactions between disease associated regions and their target gene promoters to explore disease mechanisms and finally to show that, by combining Capture Hi-C results with subsequent bioinformatics tools and experiments, it has the potential to uncover disease mechanisms and therapeutic targets.

Although largely superseded by other tools, such as HaploReg and RegulomeDB, ASSIMILATOR was the first tool developed to mine the ENCODE data and proved to be a useful tool to select potentially functional variants from GWAS associations. However the increase in the amount of epigenomic data over the last few years and the number of diverse sources, favours the use of pre-processed static datasets, such as those used by HaploReg and RegulomeDB, rather than a 'real-time', up to date, data retrieval method due to speed considerations. Additionally, with the increased numbers of cell types and experiments, researchers should consider looking at specific sub-types relevant to the disease of interest or methods to summarise experimental results into functionally relevant categories. The ChromHMM method, used to classify regions of the genome into chromatin states, offers an ideal solution to summarise experiments and allows researchers to concentrate on cell types instead. As such, prioritisation tools, such as ASSIMILATOR, should base their initial searchers on ChromHMM data and then integrate other resources, such as transcription factor binding, which as yet has not been achieved. Similarly, methods to explore genomic feature enrichment and linking environmental factors with disease causality have evolved and methods, such as Mendelian randomisation offer the potential to fully explore disease causality.

However, the current challenge, to translate GWAS findings to functional mechanisms, knowing the majority are regulatory, is to identify which gene or genes the variants act on to regulate their expression. Although this has historically been assumed to be the closest, most biologically plausible gene, we have shown, using Capture Hi-C, that this may not always be the case and variants may actually have several gene targets. Capture Hi-C has proved that gene regulation involves a complex interplay of several factors and bioinformatics has been crucial in all aspects of these experiments: in their design, analysis and interpretation. Although previous studies have used Capture Hi-C to study chromatin interactions, our Capture Hi-C experiment was the first to systematically and comprehensively investigate chromatin interactions between all regions associated with four autoimmune diseases and their target genes. Our approach to target both disease associated regions and selected promoters allowed us to explore in much greater detail, compared with a whole-genome promoter capture, how disease variants may effect gene regulation. Unexpectedly the observed interactions were much further than originally thought and although our complementary experimental design was unique to this study and offered to be a powerful approach to self-validate chromatin interactions, it did not yield as much potential as hoped. As such, our experimental design for future studies has been adapted slightly and instead of performing both capture experiments in parallel, they will be performed sequentially, with the first experiment informing the second. We believe that whilst this approach will take longer to execute it offers the most comprehensive way to validate findings and is therefore being adopted for future studies both by our group and others.

One of the main limitations of this study however, is the use of cell lines as opposed to human primary cells, either healthy controls or affected individuals. Cells lines are much more amenable than primary cells, as they overcome technical limitations with regards to cell number and availability. But despite not being alone in the use of cell lines in Capture Hi-C experiments, they do not necessarily represent a true cellular state and as such, any findings may not be comparable to an *in vivo* system. Ideally, 50 million cells are required to produce a Hi-C library to ensure complexity in the final library. For cell lines this number is easily achievable but for primary cells, less so. We have produced data showing comparable results with lower cell numbers (unpublished), which will aid the use of primary cells in future Capture Hi-C experiments. It is however imperative that future studies are performed in primary cells relevant to disease.

The final two papers not only highlight what could be achieved using Capture Hi-C, coupled with further validatory and exploratory experiments, but also the importance of cell type specificity in establishing a link between expression and genotype. They demonstrate the ability of Capture Hi-C to identify chromatin interactions which affect disease and how these findings have led to a better understanding of how the associated variants contribute to disease. Although, further validation would be required in primary cells and the direct effects of genotype specificity explored, Capture Hi-C has also demonstrated potential utility in identifying pathways involved in disease and identification of new and existing drug targets which could provide a real clinical impact and patient benefit. Recent advances in genome editing, specifically the CRISPR/Cas9 system, will allow researchers to explore the direct effect of genotype on cellular phenotype, either by directly modifying the genome or by altering gene regulation, for example the effect of the enhancer (Figure 27).

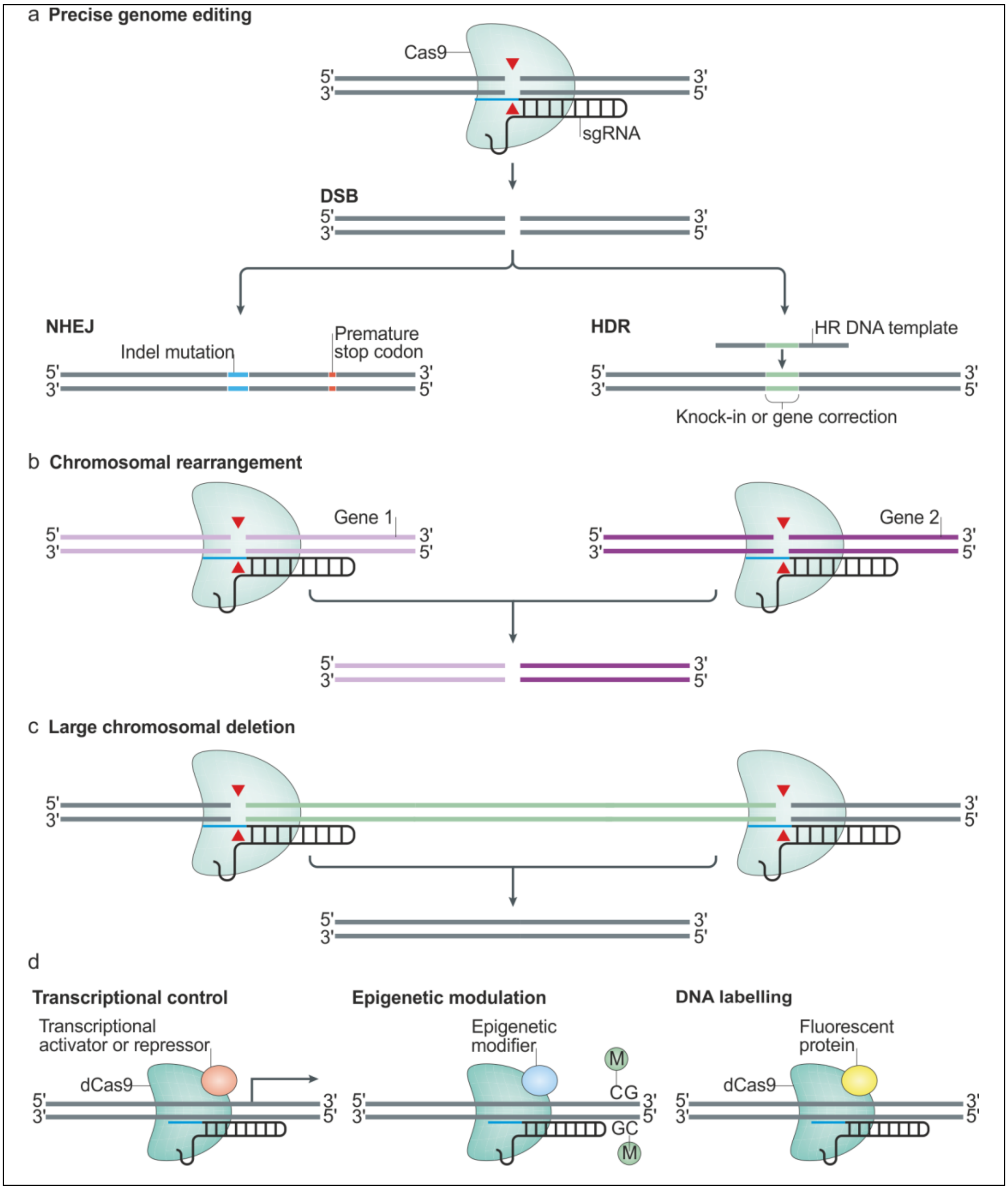


Figure 27 Applications of the CRISPR/Cas9 system

Precise genome editing (a) can be achieved by targeting a single locus. The DNA is repaired using either an error prone non-homologous end-joining (NHEJ) to produce indels or the precise homology-directed repair (HDR). Chromosomal rearrangements (b) and large chromosomal deletions (c) can be performed by targeting two different sites of the genome. The desired effect can be altered by varying the distance between the two sites. Finally, a functionally inactive or dead Cas9 (dCas9) can be fused with different functional modifier domains to induce transcriptional control, epigenetic modification or DNA labelling. Modified from Heidenreich *et al.*[171]. Adapted by permission from Macmillan Publishers Ltd: Nature Reviews Neuroscience 17:36-44, copyright 2016.

It is clear that post-GWAS, bioinformatics has informed the steps we need to take to refine, prioritise, integrate and translate GWAS findings into a complete understanding of disease. It will continue to be invaluable in future efforts to uncover these mechanisms and to inform new therapies and to better inform treatment options.

The exponential progression in the cost and throughput of data accumulation, particularly related to sequence data, has been matched by an exponential requirement to integrate robust statistical bioinformatics into genetic pipelines. Already we can see how bioinformatics has informed such diverse discoveries as the cell types important in disease, the histone marks enriched in GWAS loci, pathways important in disease, how expression changes with cell type and state, how SNPs influence the expression of certain transcript isoforms, how enhancers are defined, the genes they are linked to and the mechanism by which an implicated SNP may change expression, for example through TF binding. Bioinformatics, therefore, although still evolving, has made many positive impacts on our understanding of complex disease, and will continue to do so in the future.

6 Scientific Impact

Publication 1 was published in Bioinformatics (Oxford University Press), a leading journal in its field. In 2011 it had an impact factor of 5.468 and was ranked the number 1 journal, out of 47, in the category of mathematical and computational biology. Publication 2 was published in Arthritis and Rheumatology (Wiley-Blackwell), an official journal of the American college of rheumatology. Arthritis and Rheumatology was previously known as Arthritis and Rheumatism and in 2014 was ranked 3rd out of 31 journals in the category of rheumatology, behind Annals of Rheumatology and Nature Reviews Rheumatology with an impact factor of 7.764. Publication 3 was published in Genes and Immunity (Nature Publishing Group), a journal dedicated to functional genetics of the immune response. In 2013 it was ranked 47th out of 165 in the category of genetics and heredity and 45th out of 144 in the category of immunity, with an impact factor of 3.789. Publication 4 was published in Nature Communications, an open access journal that publishes high-quality research in biology, physics, chemistry, Earth sciences, and related areas. In 2015 it had an impact factor of 11.329 and was ranked 3rd out of 63 in the category of multidisciplinary sciences behind Nature and Science. Publication 5 was published in Genome Biology (Biomed Central Ltd), an online only, open access journal publishing outstanding research in all areas of biology and biomedicine studied from a genomic and post-genomic perspective. No impact factors or journal ranking have been released for 2016 but in 2015 it had an impact factor of 11.313 and was ranked 5th out of 161 and 7th out of 165 in the categories of biotechnology & applied microbiology and genetics & heredity respectively. Publication 6 was published in PLoS One (Public Library of Science), the world's first multidisciplinary Open Access journal, publishing reports of original research from all disciplines within science and medicine. Again, no impact factors or journal ranking have been released for 2016 but in 2015 it had an impact factor of 3.057 and was ranked 11th out of 63 in the category of multidisciplinary sciences. All impact factors and category rankings are based on Thomson Reuters™ InCites™ Journal Citation Reports® (http://jcr.incites.thomsonreuters.com).

In total, as of November 2016, the work presented here has been cited 45 times. Publication 1 has been cited seven times[31,32,103,104,172–174] and publication 2 has been cited seventeen times[175–191]. Publication 3 has been cited seven times[137,192–197]. Publication 4 has been cited fourteen times[35,36,164,175,198–207]. This publication is still relatively recent and the number of citations is expected to rise. However, the measure of online attention, Altmetric, scores this article 78, placing it in the 97th percentile of tracked articles of a similar age in all journals. Additionally, it has been viewed over 6,600 times as of November 2016. Publications 5 & 6 have only recently been published and therefore do not currently have any citations, although the number is expected to rise and publication 5 has an Altmetric score of 27 putting it in the top 5% of all tracked research outputs and the 92nd percentile of all outputs of a similar age.

In addition to the publications and citations, the work presented here has also been selected for presentation at national and international conferences. Publication 1 has been presented at a North of England Genetic Epidemiology Group meeting in Leeds, UK in November 2010. Publication 2 was selected for an oral presentation in the "Genetics" session at the British Society of Rheumatology conference in 2012 held in Glasgow, UK. Publication 3 was selected for an oral presentation in the "Genomics, genetics and epigenetics of rheumatic diseases" session at the European League Against Rheumatism (EULAR) conference in 2012 held in Berlin, Germany[208]. Publication 4 was selected for a platform (oral) presentation at the 65th American Society of Human Genetics conference held in Baltimore, MD, USA in the "Going All In: Experimental Characterization of Complex Trait Loci" session in 2015[209]. Publication 5 was selected for a poster presentation at the 65th American Society of Human Genetics conference held in Baltimore, MD, USA in 2015[210] obtaining a reviewers' choice award and oral presentations at the Target Validation using Genomics and Informatics conference held at the Wellcome Trust Genome Campus, Hinxton in 2015, the Be The Cure Functional Genomics Workshop held in London, UK in 2016 and the European Human Genetics Conference 2016 held in Barcelona, Spain in the "Complex traits" session[211]. Publication 6 was selected for a poster presentation at the European Human Genetics Conference 2016 held in Barcelona, Spain[212], obtaining a poster award. It was also selected for a platform (oral) presentation at the 66th American Society of Human Genetics conference held in Vancouver, BC, Canada in the "Chromatin Architecture, Fine Mapping, and Disease" session in 2016[213].

7 References

1. Buchwald, M. *et al.* Linkage of cystic fibrosis to the pro alpha 2(I) collagen gene, COL1A2, on chromosome 7. *Cytogenet. Cell Genet.* **41,** 234–9 (1986).

2. Gitschier, J., Drayna, D., Tuddenham, E. G., White, R. L. & Lawn, R. M. Genetic mapping and diagnosis of haemophilia A achieved through a BclI polymorphism in the factor VIII gene. *Nature* **314,** 738–740 (1985).

3. Remmers, E. F. *et al.* STAT4 and the risk of rheumatoid arthritis and systemic lupus erythematosus. *N. Engl. J. Med.* **357,** 977–86 (2007).

4. Cox, N. J. *et al.* Seven Regions of the Genome Show Evidence of Linkage to Type 1 Diabetes in a Consensus Analysis of 767 Multiplex Families. *Am. J. Hum. Genet* **69,** 820–830 (2001).

5. Siontis, K. C. M., Patsopoulos, N. a & Ioannidis, J. P. a. Replication of past candidate loci for common diseases and phenotypes in 100 genome-wide association studies. *Eur. J. Hum. Genet.* **18,** 832–7 (2010).

6. WTCCC. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* **447,** 661–78 (2007).

7. Duerr, R. H. *et al.* A genome-wide association study identifies IL23R as an inflammatory bowel disease gene. *Science* **314,** 1461–1463 (2006).

8. De Jager, P. L. *et al.* Meta-analysis of genome scans and replication identify CD6, IRF8 and TNFRSF1A as new multiple sclerosis susceptibility loci. *Nat. Genet.* **41,** 776–82 (2009).

9. Barrett, J. C. *et al.* Genome-wide association study and meta-analysis find that over 40 loci affect risk of type 1 diabetes. *Nat. Genet.* **41,** 703–7 (2009).

10. Plenge, R. M. *et al. TRAF1–C5* as a Risk Locus for Rheumatoid Arthritis — A Genomewide Study. *N. Engl. J. Med.* **357,** 1199–1209 (2007).

11. Plenge, R. M. *et al.* Two independent alleles at 6q23 associated with risk of rheumatoid arthritis. *Nat. Genet.* **39,** 1477–82 (2007).

12. Farh, K. K.-H. *et al.* Genetic and epigenetic fine mapping of causal autoimmune disease variants. *Nature* **518,** 337–43 (2015).

13. Fairfax, B. P. *et al.* Innate immune activity conditions the effect of regulatory variants upon monocyte gene expression. *Science* **343,** 1246949 (2014).

14. Ye, C. J. *et al.* Intersection of population variation and autoimmunity genetics in human T cell activation. *Science* **345,** 1254665 (2014).

15. Cortes, A. & Brown, M. A. Promise and pitfalls of the Immunochip. *Arthritis Res Ther* **13,** 101 (2011).

16. Trynka, G. *et al.* Dense genotyping identifies and localizes multiple common and rare variant association signals in celiac disease. *Nat. Genet.* **43,** 1193–201 (2011).

17. Tsoi, L. C. *et al.* Identification of 15 new psoriasis susceptibility loci highlights the role of innate immunity. *Nat. Genet.* **44,** 1341–8 (2012).

18. Cooper, J. D. *et al.* Seven newly identified loci for autoimmune thyroid

disease. *Hum. Mol. Genet.* **21,** 5202–5208 (2012).

19. Eyre, S. *et al.* High-density genetic mapping identifies new susceptibility loci for rheumatoid arthritis. *Nat. Genet.* **44,** 1336–40 (2012).

20. Liu, J. Z. *et al.* Dense fine-mapping study identifies new susceptibility loci for primary biliary cirrhosis. *Nat. Genet.* **44,** 1137–41 (2012).

21. Hinks, A. *et al.* Dense genotyping of immune-related disease regions identifies 14 new susceptibility loci for juvenile idiopathic arthritis. *Nat. Genet.* **45,** 664–9 (2013).

22. Onengut-Gumuscu, S. *et al.* Fine mapping of type 1 diabetes susceptibility loci and evidence for colocalization of causal variants with lymphoid gene enhancers. *Nat. Genet.* **47,** 381–6 (2015).

23. Cortes, A. *et al.* Identification of multiple risk variants for ankylosing spondylitis through high-density genotyping of immune-related loci. *Nat. Genet.* **45,** 730–8 (2013).

24. Faraco, J. *et al.* ImmunoChip study implicates antigen presentation to T cells in narcolepsy. *PLoS Genet.* **9,** e1003270 (2013).

25. Beecham, A. H. *et al.* Analysis of immune-related loci identifies 48 new susceptibility variants for multiple sclerosis. *Nat. Genet.* **45,** 1353–60 (2013).

26. Liu, J. Z. *et al.* Dense genotyping of immune-related disease regions identifies nine new risk loci for primary sclerosing cholangitis. *Nat. Genet.* **45,** 670–5 (2013).

27. Mayes, M. D. *et al.* Immunochip analysis identifies multiple susceptibility loci for systemic sclerosis. *Am. J. Hum. Genet.* **94,** 47–61 (2014).

28. Okada, Y. *et al.* Genetics of rheumatoid arthritis contributes to biology and drug discovery. *Nature* **506,** 376–81 (2014).

29. Martin, P., Barton, A. & Eyre, S. ASSIMILATOR: a new tool to inform selection of associated genetic variants for functional studies. *Bioinformatics* **27,** 144–6 (2011).

30. Consortium, T. E. P. The ENCODE (ENCyclopedia Of DNA Elements) Project. *Science (80-.).* **306,** 636–640 (2004).

31. Orozco, G. *et al.* Novel rheumatoid arthritis susceptibility locus at 22q12 identified in an extended UK genome-wide association study. *Arthritis Rheumatol. (Hoboken, N.J.)* **66,** 24–30 (2014).

32. Yarwood, A. *et al.* Enrichment of vitamin D response elements in RA-associated loci supports a role for vitamin D in the pathogenesis of RA. *Genes Immun.* **14,** 325–9 (2013).

33. Ramagopalan, S. V *et al.* A ChIP-seq defined genome-wide map of vitamin D receptor binding: associations with disease and evolution. *Genome Res.* **20,** 1352–60 (2010).

34. Martin, P. *et al.* Capture Hi-C reveals novel candidate genes and complex long-range interactions with related autoimmune risk loci. *Nat. Commun.* **6,** 10069 (2015).

35. McGovern, A. *et al.* Capture Hi-C identifies a novel causal gene, IL20RA, in the pan-autoimmune genetic susceptibility region 6q23. *Genome Biol.* **17,** 212 (2016).

36. Martin, P. *et al.* Identifying Causal Genes at the Multiple Sclerosis Associated Region 6q23 Using Capture Hi-C. *PLoS One* **11,** e0166923 (2016).

37. Ollier, W. & Worthington, J. Small fish in a big pond. *Br. J. Rheumatol.* **36,** 931–2 (1997).

38. Stahl, E. A. *et al.* Genome-wide association study meta-analysis identifies seven new rheumatoid arthritis risk loci. *Nat. Genet.* **42,** 508–14 (2010).

39. Raychaudhuri, S. *et al.* Common variants at CD40 and other loci confer risk of rheumatoid arthritis. *Nat. Genet.* **40,** 1216–23 (2008).

40. Raychaudhuri, S. *et al.* Genetic variants at CD28, PRDM1 and CD2/CD58 are associated with rheumatoid arthritis risk. *Nat. Genet.* **41,** 1313–8 (2009).

41. Gregersen, P. K., Silver, J. & Winchester, R. J. The shared epitope hypothesis. An approach to understanding the molecular genetics of susceptibility to rheumatoid arthritis. *Arthritis Rheum.* **30,** 1205–13 (1987).

42. Raychaudhuri, S. *et al.* Five amino acids in three HLA proteins explain most of the association between MHC and seropositive rheumatoid arthritis. *Nat. Genet.* **44,** 291–6 (2012).

43. Viatte, S. *et al.* Association of HLA-DRB1 Haplotypes With Rheumatoid Arthritis Severity, Mortality, and Treatment Response. *Jama* **313,** 1645 (2015).

44. Han, B. *et al.* Fine mapping seronegative and seropositive rheumatoid arthritis to shared and distinct HLA alleles by adjusting for the effects of heterogeneity. *Am. J. Hum. Genet.* **94,** 522–32 (2014).

45. Begovich, A. B. *et al.* A missense single-nucleotide polymorphism in a gene encoding a protein tyrosine phosphatase (PTPN22) is associated with rheumatoid arthritis. *Am J Hum Genet* **75,** 330–337 (2004).

46. Burn, G. L., Svensson, L., Sanchez-Blanco, C., Saini, M. & Cope, A. P. Why is PTPN22 a good candidate susceptibility gene for autoimmune disease? *FEBS Letters* **585,** 3689–3698 (2011).

47. Gateva, V. *et al.* A large-scale replication study identifies TNIP1, PRDM1, JAZF1, UHRF1BP1 and IL10 as risk loci for systemic lupus erythematosus. *Nat. Genet.* **41,** 1228–33 (2009).

48. Gregersen, P. K. *et al.* Risk for myasthenia gravis maps to a 151Pro→Ala change in TNIP1 and to human leukocyte antigen-B*08. *Ann. Neurol.* **72,** 927–935 (2012).

49. Barrett, J. C. *et al.* Genome-wide association defines more than thirty distinct susceptibility loci for Crohn's disease. *Nat Genet.* **40,** 955–962 (2009).

50. Hinks, A. *et al.* 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,** 1694–1699 (2005).

51. Bottini, N. *et al.* A functional variant of lymphoid tyrosine phosphatase is associated with type I diabetes. *Nat. Genet.* **36,** 337–8 (2004).

52. Vang, T. *et al.* Autoimmune-associated lymphoid tyrosine phosphatase is a gain-of-function variant. *Nat. Genet.* **37,** 1317–1319 (2005).

53. Rieck, M. *et al.* Genetic variation in PTPN22 corresponds to altered function of T and B lymphocytes. *J. Immunol. (Baltimore, Md 1950)* **179,** 4704–4710 (2007).

54. Zikherman, J. *et al.* PTPN22 deficiency cooperates with the CD45 E613R allele to break tolerance on a non-autoimmune background. *J. Immunol.* **182,** 4093–4106 (2009).

55. Lefvert, A. K. *et al.* PTPN22 R620W promotes production of anti-AChR autoantibodies and IL-2 in myasthenia gravis. *J. Neuroimmunol.* **197,** 110–113 (2008).

56. Zhang, J. *et al.* The autoimmune disease-associated PTPN22 variant promotes calpain-mediated Lyp/Pep degradation associated with lymphocyte and dendritic cell hyperresponsiveness. *Nat. Genet.* **43,** 902–907 (2011).

57. Seidl, C. *et al.* CTLA4 codon 17 dimorphism in patients with rheumatoid arthritis. *Tissue Antigens* **51,** 62–6 (1998).

58. Gregersen, P. K. *et al.* REL, encoding a member of the NF-kappaB family of transcription factors, is a newly defined risk locus for rheumatoid arthritis. *Nat. Genet.* **41,** 820–3 (2009).

59. Thomson, W. *et al.* Rheumatoid arthritis association at 6q23. *Nat. Genet.* **39,** 1431–1433 (2007).

60. Zhernakova, A. *et al.* Meta-analysis of genome-wide association studies in celiac disease and rheumatoid arthritis identifies fourteen non-HLA shared loci. *PLoS Genet.* **7,** e1002004 (2011).

61. Suzuki, A. *et al.* Functional haplotypes of PADI4, encoding citrullinating enzyme peptidylarginine deiminase 4, are associated with rheumatoid arthritis. *Nat. Genet.* **10,** 520–527 (2003).

62. Viatte, S., Plant, D. & Raychaudhuri, S. Genetics and epigenetics of rheumatoid arthritis. *Nat. Rev. Rheumatol.* **9,** 141–53 (2013).

63. Terao, C., Raychaudhuri, S. & Gregersen, P. K. Recent Advances in Defining the Genetic Basis of Rheumatoid Arthritis. *Annu. Rev. Genomics Hum. Genet.* **17,** annurev-genom-090314-045919 (2016).

64. Todd, J. a *et al.* Robust associations of four new chromosome regions from genome-wide analyses of type 1 diabetes. *Nat. Genet.* **39,** 857–64 (2007).

65. Bradfield, J. P. *et al.* A genome-wide meta-analysis of six type 1 diabetes cohorts identifies multiple associated loci. *PLoS Genet.* **7,** e1002293 (2011).

66. Franke, A. *et al.* Genome-wide meta-analysis increases to 71 the number of confirmed Crohn's disease susceptibility loci. *Nat. Genet.* **42,** 1118–25 (2010).

67. Jostins, L. *et al.* Host-microbe interactions have shaped the genetic

architecture of inflammatory bowel disease. *Nature* **491,** 119–24 (2012).

68. Esparza-Gordillo, J. *et al.* A functional IL-6 receptor (IL6R) variant is a risk factor for persistent atopic dermatitis. *J. Allergy Clin. Immunol.* **132,** 371–377 (2013).

69. Bowes, J. *et al.* Dense genotyping of immune-related susceptibility loci reveals new insights into the genetics of psoriatic arthritis. *Nat. Commun.* **6,** 6046 (2015).

70. Trynka, G. *et al.* Disentangling the Effects of Colocalizing Genomic Annotations to Functionally Prioritize Non-coding Variants within Complex-Trait Loci. *Am. J. Hum. Genet.* **97,** 139–52 (2015).

71. FitzGerald, O., Haroon, M., Giles, J. T. & Winchester, R. Concepts of pathogenesis in psoriatic arthritis: genotype determines clinical phenotype. *Arthritis Res. Ther.* **17,** 115 (2015).

72. Menon, B. *et al.* Interleukin-17+CD8+ T cells are enriched in the joints of patients with psoriatic arthritis and correlate with disease activity and joint damage progression. *Arthritis Rheumatol.* **66,** 1272–1281 (2014).

73. Berglin, E. *et al.* Radiological outcome in rheumatoid arthritis is predicted by presence of antibodies against cyclic citrullinated peptide before and at disease onset, and by IgA-RF at disease onset. *Ann. Rheum. Dis.* **65,** 453–458 (2006).

74. Bukhari, M. *et al.* The performance of anti-cyclic citrullinated peptide antibodies in predicting the severity of radiologic damage in inflammatory polyarthritis: Results from the Norfolk Arthritis Register. *Arthritis Rheum.* **56,** 2929–2935 (2007).

75. Van Der Woude, D. *et al.* Quantitative heritability of anti-citrullinated protein antibody-positive and anti-citrullinated protein antibody-negative rheumatoid arthritis. *Arthritis Rheum.* **60,** 916–923 (2009).

76. Frisell, T. *et al.* Familial risks and heritability of rheumatoid arthritis: Role of rheumatoid factor/anti-citrullinated protein antibody status, number and type of affected relatives, sex, and age. *Arthritis Rheum.* **65,** 2773–2782 (2013).

77. Viatte, S. *et al.* Replication of Associations of Genetic Loci Outside the HLA Region With Susceptibility to Anti-Cyclic Citrullinated Peptide-Negative Rheumatoid Arthritis. *Arthritis Rheumatol. (Hoboken, N.J.)* **68,** 1603–13 (2016).

78. Viatte, S. *et al.* Genetic markers of rheumatoid arthritis susceptibility in anti-citrullinated peptide antibody negative patients. *Ann. Rheum. Dis.* **71,** 1984–90 (2012).

79. Breedveld, F. C. & Verweij, C. L. T cells in rheumatoid arthritis. *Br. J. Rheumatol.* **36,** 617–619 (1997).

80. Stanford, S. M. & Bottini, N. PTPN22: the archetypal non-HLA autoimmunity gene. *Nat. Rev. Rheumatol.* **10,** 602–611 (2014).

81. Maine, C. J. *et al.* PTPN22 alters the development of regulatory T cells in the thymus. *J. Immunol.* **188,** 5267–75 (2012).

82. Brownlie, R. J. *et al.* Lack of the phosphatase PTPN22 increases adhesion of murine regulatory T cells to improve their immunosuppressive function. *Sci. Signal.* **5,** ra87 (2012).

83. Kochi, Y. *et al.* A regulatory variant in CCR6 is associated with rheumatoid arthritis susceptibility. *Nat. Genet.* **42,** 515–519 (2010).

84. Dendrou, C. A. *et al.* Cell-specific protein phenotypes for the autoimmune locus IL2RA using a genotype-selectable human bioresource. *Nat. Genet.* **41,** 1011–5 (2009).

85. Eleftherohorinou, H. *et al.* Pathway Analysis of GWAS Provides New Insights into Genetic Susceptibility to 3 Inflammatory Diseases. *PLoS One* **4,** e8068 (2009).

86. Richardson, B. *et al.* Evidence for impaired T cell DNA methylation in systemic lupus erythematosus and rheumatoid arthritis. *Arthritis Rheum.* **33,** 1665–1673 (1990).

87. Trynka, G. *et al.* Chromatin marks identify critical cell types for fine mapping complex trait variants. *Nat. Genet.* **45,** 124–30 (2013).

88. Hu, X. *et al.* Integrating autoimmune risk loci with gene-expression data identifies specific pathogenic immune cell subsets. *Am. J. Hum. Genet.* **89,** 496–506 (2011).

89. Rudd, C. E. & Schneider, H. Unifying concepts in CD28, ICOS and CTLA4 co-receptor signalling. *Nat. Rev. Immunol.* **3,** 544–556 (2003).

90. Lund, F. E., Garvy, B. A., Randall, T. D. & Harris, D. P. Regulatory roles for cytokine-producing B cells in infection and autoimmune disease. *Curr. Dir. Autoimmun.* **8,** 25–54 (2005).

91. Clark, E. A. & Ledbetter, J. A. How does B cell depletion therapy work, and how can it be improved? *Ann. Rheum. Dis.* **64 Suppl 4,** iv77-80 (2005).

92. Bugatti, S., Vitolo, B., Caporali, R., Montecucco, C. & Manzo, A. B cells in rheumatoid arthritis: From pathogenic players to disease biomarkers. *Biomed Res. Int.* **2014,** 681678 (2014).

93. Shaw, T., Quan, J. & Totoritis, M. C. B cell therapy for rheumatoid arthritis: the rituximab (anti-CD20) experience. *Ann. Rheum. Dis.* **62 Suppl 2,** ii55-9 (2003).

94. Owen, S. A. *et al.* Genetic polymorphisms in key methotrexate pathway genes are associated with response to treatment in rheumatoid arthritis patients. *Pharmacogenomics J.* **13,** 227–34 (2013).

95. Blumenauer, B. *et al.* Infliximab for the treatment of rheumatoid arthritis. *Cochrane database Syst. Rev.* CD003785 (2002). doi:10.1002/14651858.CD003785

96. Martin, P. *et al.* Chromatin Interactions Reveal Novel Gene Targets for Drug Repositioning in Rheumatic Diseases. *Arthritis Rheumatol.* **68,** (Suppl 10), Abstract #76 (2016).

97. Ecker, J. R. *et al.* Genomics: ENCODE explained. *Nature* **489,** 52–5 (2012).

98. Annunziato, A. DNA Packaging: Nucleosomes and Chromatin. *Nat. Educ.* **1,** 26 (2008).

99. Birney, E. *et al.* Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature* **447,** 799–816 (2007).

100. Kent, W. J. *et al.* The human genome browser at UCSC. *Genome Res.* **12,** 996–1006 (2002).

101. Shlyueva, D., Stampfel, G. & Stark, A. Transcriptional enhancers: from properties to genome-wide predictions. *Nat. Rev. Genet.* **15,** 272–86 (2014).

102. Pomerantz, M. M. *et al.* The 8q24 cancer risk variant rs6983267 shows long-range interaction with MYC in colorectal cancer. *Nat. Genet.* **41,** 882–4 (2009).

103. Cobb, J. *et al.* Genome-wide data reveal novel genes for methotrexate response in a large cohort of juvenile idiopathic arthritis cases. *Pharmacogenomics J.* **14,** 356–64 (2014).

104. Warren, R. B. *et al.* A systematic investigation of confirmed autoimmune loci in early-onset psoriasis reveals an association with IL2/IL21. *Br. J. Dermatol.* **164,** 660–4 (2011).

105. Ward, L. D. & Kellis, M. HaploReg: a resource for exploring chromatin states, conservation, and regulatory motif alterations within sets of genetically linked variants. *Nucleic Acids Res.* **40,** D930-4 (2012).

106. Boyle, A. P. *et al.* Annotation of functional variation in personal genomes using RegulomeDB. *Genome Res.* **22,** 1790–7 (2012).

107. Matys, V. *et al.* TRANSFAC(R) and its module TRANSCompel(R): transcriptional gene regulation in eukaryotes. *Nucl. Acids Res.* **34,** D108-110 (2006).

108. Portales-Casamar, E. *et al.* JASPAR 2010: The greatly expanded open-access database of transcription factor binding profiles. *Nucleic Acids Res.* **38,** D105-10 (2009).

109. Auton, A. *et al.* A global reference for human genetic variation. *Nature* **526,** 68–74 (2015).

110. Bernstein, B. E. *et al.* The NIH Roadmap Epigenomics Mapping Consortium. *Nat. Biotechnol.* **28,** 1045–8 (2010).

111. Lonsdale, J. *et al.* The Genotype-Tissue Expression (GTEx) project. *Nat. Genet.* **45,** 580–5 (2013).

112. Ritchie, G. R. S., Dunham, I., Zeggini, E. & Flicek, P. Functional annotation of noncoding sequence variants. *Nat. Methods* **11,** 294–6 (2014).

113. Ryan, N. M., Morris, S. W., Porteous, D. J., Taylor, M. S. & Evans, K. L. SuRFing the genomics wave: an R package for prioritising SNPs by functionality. *Genome Med.* **6,** 79 (2014).

114. Gulko, B., Hubisz, M. J., Gronau, I. & Siepel, A. A method for calculating probabilities of fitness consequences for point mutations across the human genome. *Nat. Publ. Gr.* **47,** 276–283 (2015).

115. Kircher, M. *et al.* A general framework for estimating the relative pathogenicity of human genetic variants. *Nat. Genet.* **46,** 310–315 (2014).

116. McLaren, W. *et al.* The Ensembl Variant Effect Predictor. *Genome Biol.* **17,** 122 (2016).

117. Cooper, G., Kircher, M., Witten, D. & Shendure, J. Probabilities of Fitness Consequences for Point Mutations Across the Human Genome. (2014). Available at: https://haldanessieve.org/2014/10/23/thoughts-on-probabilities-of-fitness-consequences-for-point-mutations-across-the-human-genome/. (Accessed: 22nd August 2016)

118. Ernst, J. & Kellis, M. ChromHMM: automating chromatin-state discovery and characterization. *Nat. Methods* **9,** 215–6 (2012).

119. Ernst, J. & Kellis, M. Large-scale imputation of epigenomic datasets for systematic annotation of diverse human tissues. *Nat. Biotechnol.* **33,** 364–76 (2015).

120. Hunter, D. J. Gene-environment interactions in human diseases. *Nat. Rev. Genet.* **6,** 287–98 (2005).

121. Sturm, R. A. Skin colour and skin cancer - MC1R, the genetic link. *Melanoma Res.* **12,** 405–16 (2002).

122. Thomas, D. Methods for investigating gene-environment interactions in candidate pathway and genome-wide association studies. *Annu. Rev. Public Health* **31,** 21–36 (2010).

123. Holick, M. F. Vitamin D deficiency. *N. Engl. J. Med.* **357,** 266–81 (2007).

124. Kato, S. The function of vitamin D receptor in vitamin D action. *J. Biochem.* **127,** 717–22 (2000).

125. Kostoglou-Athanassiou, I., Athanassiou, P., Lyraki, A., Raftakis, I. & Antoniadis, C. Vitamin D and rheumatoid arthritis. *Ther. Adv. Endocrinol. Metab.* **3,** 181–7 (2012).

126. Correale, J., Ysrraelit, M. C. & Gaitán, M. I. Immunomodulatory effects of Vitamin D in multiple sclerosis. *Brain* **132,** 1146–60 (2009).

127. Jirapongsananuruk, O., Melamed, I. & Leung, D. Y. Additive immunosuppressive effects of 1,25-dihydroxyvitamin D3 and corticosteroids on TH1, but not TH2, responses. *J. Allergy Clin. Immunol.* **106,** 981–5 (2000).

128. Bartels, L. E., Hvas, C. L., Agnholt, J., Dahlerup, J. F. & Agger, R. Human dendritic cell antigen presentation and chemotaxis are inhibited by intrinsic 25-hydroxy vitamin D activation. *Int. Immunopharmacol.* **10,** 922–8 (2010).

129. Song, G. G., Bae, S.-C. & Lee, Y. H. Association between vitamin D intake and the risk of rheumatoid arthritis: a meta-analysis. *Clin. Rheumatol.* **31,** 1733–9 (2012).

130. Dahlquist, G. Vitamin D supplement in early childhood and risk for Type I (insulin- dependent) diabetes mellitus. *Diabetologia* **42,** 51–54 (1999).

131. Stene, L. C., Ulriksen, J., Magnus, P. & Joner, G. Use of cod liver oil during pregnancy associated with lower risk of Type I diabetes in the offspring.

Diabetologia **43,** 1093–8 (2000).

132. van der Mei, I. A. F. *et al.* Past exposure to sun, skin phenotype, and risk of multiple sclerosis: case-control study. *BMJ* **327,** 316 (2003).

133. Munger, K. L. *et al.* Vitamin D intake and incidence of multiple sclerosis. *Neurology* **62,** 60–5 (2004).

134. Merlino, L. A. *et al.* Vitamin D intake is inversely associated with rheumatoid arthritis: results from the Iowa Women's Health Study. *Arthritis Rheum.* **50,** 72–7 (2004).

135. Wang, T. J. *et al.* Common genetic determinants of vitamin D insufficiency: a genome-wide association study. *Lancet (London, England)* **376,** 180–8 (2010).

136. Lee, Y. H., Bae, S.-C., Choi, S. J., Ji, J. D. & Song, G. G. Associations between vitamin D receptor polymorphisms and susceptibility to rheumatoid arthritis and systemic lupus erythematosus: a meta-analysis. *Mol. Biol. Rep.* **38,** 3643–51 (2011).

137. Viatte, S. *et al.* The role of genetic polymorphisms regulating vitamin D levels in rheumatoid arthritis outcome: a Mendelian randomisation approach. *Ann. Rheum. Dis.* **73,** 1430–3 (2014).

138. Gray, R. & Wheatley, K. How to avoid bias when comparing bone marrow transplantation with chemotherapy. *Bone Marrow Transplant.* **7 Suppl 3,** 9–12 (1991).

139. Katan, M. B. Apolipoprotein E isoforms, serum cholesterol, and cancer. *International Journal of Epidemiology* **33,** 9 (2004).

140. Robinson, P. C., Choi, H. K., Do, R. & Merriman, T. R. Insight into rheumatological cause and effect through the use of Mendelian randomization. *Nat. Rev. Rheumatol.* **12,** 486–96 (2016).

141. Pierce, B. L., Ahsan, H. & Vanderweele, T. J. Power and instrument strength requirements for Mendelian randomization studies using multiple genetic variants. *Int. J. Epidemiol.* **40,** 740–752 (2011).

142. C Reactive Protein Coronary Heart Disease Genetics Collaboration (CCGC) *et al.* Association between C reactive protein and coronary heart disease: mendelian randomisation analysis based on individual participant data. *BMJ* **342,** d548 (2011).

143. Vimaleswaran, K. S. *et al.* Causal Relationship between Obesity and Vitamin D Status: Bi-Directional Mendelian Randomization Analysis of Multiple Cohorts. *PLoS Med.* **10,** e1001383 (2013).

144. Freitag, D. *et al.* Cardiometabolic effects of genetic upregulation of the interleukin 1 receptor antagonist: A Mendelian randomisation analysis. *Lancet Diabetes Endocrinol.* **3,** 243–253 (2015).

145. Musunuru, K. *et al.* From noncoding variant to phenotype via SORT1 at the 1p13 cholesterol locus. *Nature* **466,** 714–9 (2010).

146. Davison, L. J. *et al.* Long-range DNA looping and gene expression analyses identify DEXI as an autoimmune disease candidate gene. *Hum. Mol. Genet.*

21, 322–333 (2012).

147. Sanyal, A., Lajoie, B. R., Jain, G. & Dekker, J. The long-range interaction landscape of gene promoters. *Nature* **489,** 109–13 (2012).

148. Velagaleti, G. V. N. *et al.* Position effects due to chromosome breakpoints that map approximately 900 Kb upstream and approximately 1.3 Mb downstream of SOX9 in two patients with campomelic dysplasia. *Am. J. Hum. Genet.* **76,** 652–62 (2005).

149. Herranz, D. *et al.* A NOTCH1-driven MYC enhancer promotes T cell development, transformation and acute lymphoblastic leukemia. *Nat. Med.* **20,** 1130–7 (2014).

150. Lettice, L. A. *et al.* A long-range Shh enhancer regulates expression in the developing limb and fin and is associated with preaxial polydactyly. *Hum. Mol. Genet.* **12,** 1725–35 (2003).

151. Kagey, M. H. *et al.* Mediator and cohesin connect gene expression and chromatin architecture. *Nature* **467,** 430–5 (2010).

152. Tolhuis, B., Palstra, R. J., Splinter, E., Grosveld, F. & De Laat, W. Looping and interaction between hypersensitive sites in the active β-globin locus. *Mol. Cell* **10,** 1453–1465 (2002).

153. Stadhouders, R. *et al.* Dynamic long-range chromatin interactions control Myb proto-oncogene transcription during erythroid development. *EMBO J.* **31,** 986–99 (2012).

154. Dixon, J. R. *et al.* Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature* **485,** 376–80 (2012).

155. Lupiáñez, D. G. *et al.* Disruptions of topological chromatin domains cause pathogenic rewiring of gene-enhancer interactions. *Cell* **161,** 1012–25 (2015).

156. Schoenfelder, S. *et al.* The pluripotent regulatory circuitry connecting promoters to their long-range interacting elements. *Genome Res.* **25,** 582–597 (2015).

157. Gnirke, A. *et al.* Solution hybrid selection with ultra-long oligonucleotides for massively parallel targeted sequencing. *Nat. Biotechnol.* **27,** 182–9 (2009).

158. Jäger, R. *et al.* Capture Hi-C identifies the chromatin interactome of colorectal cancer risk loci. *Nat. Commun.* **6,** 6178 (2015).

159. Dryden, N. H. *et al.* Unbiased analysis of potential targets of breast cancer susceptibility loci by Capture Hi-C. *Genome Res.* **24,** 1854–68 (2014).

160. Mifsud, B. *et al.* Mapping long-range promoter contacts in human cells with high-resolution capture Hi-C. *Nat. Genet.* **47,** 598–606 (2015).

161. Kolovos, P. *et al.* Targeted Chromatin Capture (T2C): a novel high resolution high throughput method to detect genomic interactions and regulatory elements. *Epigenetics Chromatin* **7,** 10 (2014).

162. Hughes, J. R. *et al.* Analysis of hundreds of cis-regulatory landscapes at high resolution in a single, high-throughput experiment. *Nat. Genet.* **46,** 205–12 (2014).

163. Rao, S. S. P. *et al.* A 3D Map of the Human Genome at Kilobase Resolution Reveals Principles of Chromatin Looping. *Cell* **159,** 1665–80 (2014).

164. Javierre, B. M. *et al.* Lineage-Specific Genome Architecture Links Enhancers and Non-coding Disease Variants to Target Gene Promoters. *Cell* **167,** 1369–1384.e19 (2016).

165. Gottlieb, A. B., Krueger, J. G., Sandberg Lundblad, M., Göthberg, M. & Skolnick, B. E. First-In-Human, Phase 1, Randomized, Dose-Escalation Trial with Recombinant Anti-IL-20 Monoclonal Antibody in Patients with Psoriasis. *PLoS One* **10,** e0134703 (2015).

166. Šenolt, L. *et al.* Efficacy and Safety of Anti-Interleukin-20 Monoclonal Antibody in Patients With Rheumatoid Arthritis: A Randomized Phase IIa Trial. *Arthritis Rheumatol. (Hoboken, N.J.)* **67,** 1438–48 (2015).

167. Lander, E. S. *et al.* Initial sequencing and analysis of the human genome. *Nature* **409,** 860–921 (2001).

168. Venter, J. C. *et al.* The sequence of the human genome. *Science* **291,** 1304–51 (2001).

169. Macosko, E. Z. *et al.* Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets. *Cell* **161,** 1202–14 (2015).

170. Pollen, A. A. *et al.* Low-coverage single-cell mRNA sequencing reveals cellular heterogeneity and activated signaling pathways in developing cerebral cortex. *Nat. Biotechnol.* **32,** 1053–8 (2014).

171. Heidenreich, M. & Zhang, F. Applications of CRISPR-Cas systems in neuroscience. *Nat. Rev. Neurosci.* **17,** 36–44 (2016).

172. Hébert, H. L. *et al.* Identification of loci associated with late-onset psoriasis using dense genotyping of immune-related regions. *Br. J. Dermatol.* **172,** 933–9 (2015).

173. Kennedy, R. B. & Poland, G. A. The top five 'game changers' in vaccinology: toward rational and directed vaccine development. *OMICS* **15,** 533–7 (2011).

174. McAllister, K., Eyre, S. & Orozco, G. Genetics of rheumatoid arthritis: GWAS and beyond. *Open access Rheumatol. Res. Rev.* **3,** 31–46 (2011).

175. Kim, K., Bang, S.-Y., Lee, H.-S. & Bae, S.-C. Update on the genetic architecture of rheumatoid arthritis. *Nat. Rev. Rheumatol.* (2016). doi:10.1038/nrrheum.2016.176

176. Lu, Q. Unmet needs in autoimmunity and potential new tools. *Clin. Rev. Allergy Immunol.* **47,** 111–8 (2014).

177. Picerno, V. *et al.* One year in review: the pathogenesis of rheumatoid arthritis. *Clin. Exp. Rheumatol.* **33,** 551–8 (2015).

178. Canet, L. M. *et al.* Genetic variants within the TNFRSF1B gene and susceptibility to rheumatoid arthritis and response to anti-TNF drugs: a multicenter study. *Pharmacogenet. Genomics* **25,** 323–33 (2015).

179. Cherednichenko, A. A. *et al.* Prevalence of gene polymorphisms associated with immune disorders in populations of Northern Eurasia. *Mol. Biol.* **49,** 881–

889 (2015).

180. Goris, A. *et al.* Genetic variants are major determinants of CSF antibody levels in multiple sclerosis. *Brain* **138,** 632–43 (2015).

181. Goulielmos, G. N. *et al.* Genetic data: The new challenge of personalized medicine, insights for rheumatoid arthritis patients. *Gene* **583,** 90–101 (2016).

182. Li, S. *et al.* RBP-J imposes a requirement for ITAM-mediated costimulation of osteoclastogenesis. *J. Clin. Invest.* **124,** 5057–73 (2014).

183. Milani, L., Leitsalu, L. & Metspalu, A. An epidemiological perspective of personalized medicine: the Estonian experience. *J. Intern. Med.* **277,** 188–200 (2015).

184. Miller, C. H. *et al.* RBP-J-Regulated miR-182 Promotes TNF-α-Induced Osteoclastogenesis. *J. Immunol.* **196,** 4977–86 (2016).

185. Nabi, G. *et al.* Meta-analysis reveals PTPN22 1858C/T polymorphism confers susceptibility to rheumatoid arthritis in Caucasian but not in Asian population. *Autoimmunity* **49,** 197–210 (2016).

186. Oliver, J., Plant, D., Webster, A. P. & Barton, A. Genetic and genomic markers of anti-TNF treatment response in rheumatoid arthritis. *Biomark. Med.* **9,** 499–512 (2015).

187. Richard, A. C. *et al.* The TNF-family cytokine TL1A: from lymphocyte costimulator to disease co-conspirator. *J. Leukoc. Biol.* **98,** 333–345 (2015).

188. Tian, J. *et al.* SF3A1 and pancreatic cancer: new evidence for the association of the spliceosome and cancer. *Oncotarget* **6,** 37750–7 (2015).

189. Ward-Kavanagh, L. K., Lin, W. W., Šedý, J. R. & Ware, C. F. The TNF Receptor Superfamily in Co-stimulating and Co-inhibitory Responses. *Immunity* **44,** 1005–19 (2016).

190. Yarwood, A., Huizinga, T. W. J. & Worthington, J. The genetics of rheumatoid arthritis: risk and protection in different stages of the evolution of RA. *Rheumatology (Oxford).* **55,** 199–209 (2016).

191. Chung, I.-M., Ketharnathan, S., Thiruvengadam, M. & Rajakumar, G. Rheumatoid Arthritis: The Stride from Research to Clinical Practice. *Int. J. Mol. Sci.* **17,** (2016).

192. Arthritis, R. Hypotheses and Emerging Facts : the Vitamin D Receptor. *Int. J. Orthop.* **1,** 1–8 (2014).

193. Goldsmith, J. R. Vitamin D as an Immunomodulator: Risks with Deficiencies and Benefits of Supplementation. *Healthc. (Basel, Switzerland)* **3,** 219–32 (2015).

194. Jeffery, L. E., Raza, K. & Hewison, M. Vitamin D in rheumatoid arthritis—towards clinical application. *Nat. Rev. Rheumatol.* **12,** 201–210 (2015).

195. Boef, A. G. C., Dekkers, O. M. & le Cessie, S. Mendelian randomization studies: a review of the approaches used and the quality of reporting. *Int. J. Epidemiol.* **44,** 496–511 (2015).

196. Zhou, J. *et al.* Exploration of the serum metabolite signature in patients with rheumatoid arthritis using gas chromatography-mass spectrometry. *J. Pharm. Biomed. Anal.* **127,** 60–7 (2016).

197. Zhang, Y.-H. *et al.* The Use of Gene Ontology Term and KEGG Pathway Enrichment for Analysis of Drug Half-Life. *PLoS One* **11,** e0165496 (2016).

198. Spurrell, C. H. *et al.* The Ties That Bind: Mapping the Dynamic Enhancer-Promoter Interactome. *Cell* **167,** 1163–1166 (2010).

199. Hahn, S. & Kim, D. Relationship between spatial organization and biological function, analyzed using gene ontology and chromosome conformation capture of human and fission yeast genomes. *Genes and Genomics* **38,** 693–705 (2016).

200. Pancaldi, V. *et al.* Integrating epigenomic data and 3D genomic structure with a new measure of chromatin assortativity. *Genome Biol.* **17,** 152 (2016).

201. Withoff, S., Li, Y., Jonkers, I. & Wijmenga, C. Understanding Celiac Disease by Genomics. *Trends Genet.* **32,** 295–308 (2016).

202. Schmitt, A. D., Hu, M. & Ren, B. Genome-wide mapping and analysis of chromosome architecture. *Nat. Rev. Mol. Cell Biol.* (2016). doi:10.1038/nrm.2016.104

203. Sung, M. K., Jang, J., Lee, K. S., Ghim, C.-M. & Choi, J. K. Selected heterozygosity at cis -regulatory sequences increases the expression homogeneity of a cell population in humans. *Genome Biol.* **17,** 164 (2016).

204. Giorgetti, L. & Heard, E. Closing the loop: 3C versus DNA FISH. *Genome Biol.* **17,** 215 (2016).

205. Li, R., Liu, Y., Li, T. & Li, C. 3Disease Browser: A Web server for integrating 3D genome and disease-associated chromosome rearrangement data. *Sci. Rep.* **6,** 34651 (2016).

206. Yarwood, A., Eyre, S. & Worthington, J. Genetic susceptibility to rheumatoid arthritis and its implications for novel drug discovery. *Expert Opin. Drug Discov.* **11,** 805–13 (2016).

207. Kyttaris, V. C., Katsiari, C. G., Juang, Y.-T. & Tsokos, G. C. New insights into the pathogenesis of systemic lupus erythematosus. *Curr. Rheumatol. Rep.* **7,** 469–75 (2005).

208. Yarwood, A. *et al.* OP0210 Enrichment of vitamin D response elements in RA associated loci, suggests a role for vitamin D in the pathogenesis of RA. *Ann. Rheum. Dis.* **71,** 126.3-127 (2013).

209. Martin, P. *et al.* Chromosome interaction analysis of risk loci in related autoimmune diseases reveals complex, long-range promoter interactions implicating novel candidate genes. in *65th Annual Meeting of The American Society of Human Genetics* Abstract #70 (2015).

210. Orozco, G. *et al.* Capture Hi-C reveals a novel causal gene, IL20RA, in the pan-auto-immune genetic susceptibility region 6q23. in *65th Annual Meeting of The American Society of Human Genetics* Abstract #831T (2015).

211. McGovern, A. *et al.* Capture Hi-C reveals a novel causal gene, IL20RA, in the

pan-autoimmune genetic susceptibility region 6q23. *Eur. J. Hum. Genet.* **24,** 22 (2016).

212. Martin, P. *et al.* Capture Hi-C identifies compelling candidate causal genes and enhancers for multiple sclerosis in the 6q23 region. *Eur. J. Hum. Genet.* **24,** 360 (2016).

213. Martin, P. *et al.* Capture Hi-C identifies compelling candidate causal genes and enhancers for multiple sclerosis in the 6q23 region. in *66th Annual Meeting of The American Society of Human Genetics* Abstract #202 (2016).

Publication 1: ASSIMILATOR: a new tool to inform selection of associated genetic variants for functional studies.

BIOINFORMATICS APPLICATIONS NOTE

Vol. 27 no. 1 2011, pages 144–146
doi:10.1093/bioinformatics/btq611

Data and text mining

ASSIMILATOR: a new tool to inform selection of associated genetic variants for functional studies

Paul Martin*, Anne Barton and Stephen Eyre

Arthritis Research UK Epidemiology Unit, Manchester Academic Health Science Centre, The University of Manchester, Stopford Building, Oxford Road, Manchester, M13 9PT, UK

Associate Editor: Jonathan Wren

ABSTRACT

Motivation: Fine-mapping experiments from genome-wide association studies (GWAS) are underway for many complex diseases. These are likely to identify a number of putative causal variants, which cannot be separated further in terms of strength of genetic association due to linkage disequilibrium. The challenge will be selecting which variant to prioritize for subsequent expensive functional studies. A wealth of functional information generated from wet lab experiments now exists but cannot be easily interrogated by the user. Here, we describe a program designed to quickly assimilate this data called ASSIMILATOR and validate the method by interrogating two regions to show its effectiveness.

Availability: http://www.medicine.manchester.ac.uk/musculoskeletal/research/arc/genetics/bioinformatics/assimilator/.

Contact: paul.martin-2@manchester.ac.uk

Received on June 4, 2010; revised on October 26, 2010; accepted on October 27, 2010

1 INTRODUCTION

Genome-wide association studies (GWAS) have been enormously successful in identifying regions associated with a variety of complex traits and diseases. Fine-mapping studies are underway for many of these disorders and are likely to identify a number of putative causal variants. The challenge then will be to prioritize which variants to select for the expensive functional studies required to fully translate how these variants affect risk. In many cases, it is expected that the likely causal variants will be single nucleotide polymorphism (SNP) markers that are in complete linkage disequilibrium and which cannot be prioritized further based on genetic evidence alone. SNPs within genes which affect the resulting protein or lie in a regulatory region would be obvious candidates for functional studies but, in many complex diseases, the causal SNPs identified to date map to intergenic, non-coding regions and it is more challenging to prioritize these based on likely function (Barton *et al.*, 2008; Thomson *et al.*, 2007; Wellcome Trust Case Control Consortium, 2007).

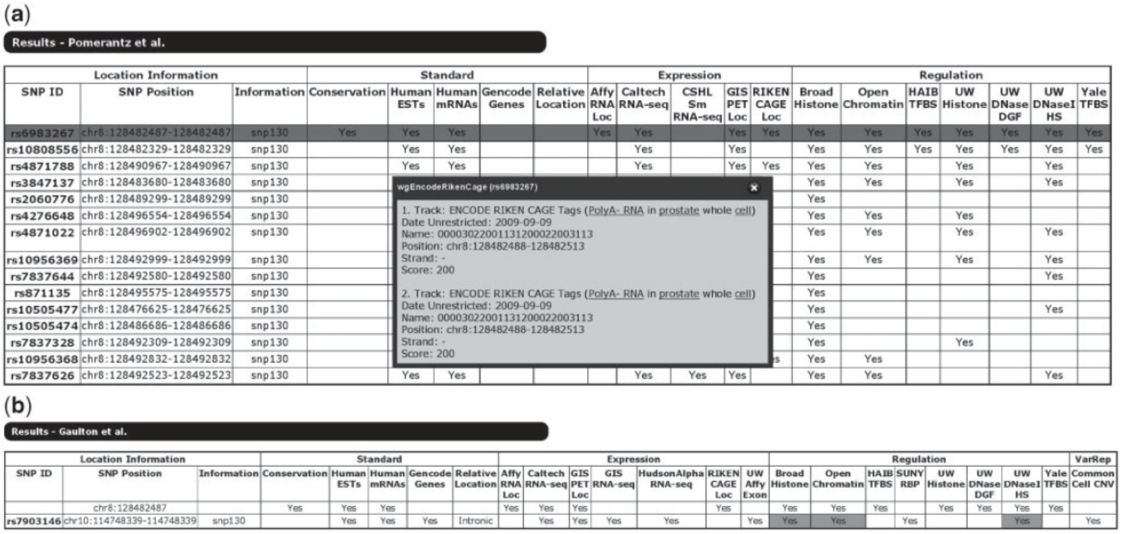
There is now a wealth of information available from the ENCyclopaedia Of DNA Elements (ENCODE) international consortium (Birney *et al.*, 2007; ENCODE Project Consortium 2004) hosted by the University of California Santa Cruz (UCSC) through their Genome Browser (Kent *et al.*, 2002). These data have been generated from wet lab experiments including Chromatin ImmunoPrecipitation Sequencing (ChIP-Seq), DNase hypersensitivity and histone modification studies, and thus may provide better evidence of putative function compared with predictive algorithms used previously to infer function at a locus. An enormous amount of data is available including studies in different cell lines and different cell compartments, but currently these sites cannot be easily interrogated by the user simultaneously. Other potential resources for prioritizing SNPs for functional studies are now becoming more widely available and include eQTL studies and programs which predict likely effects of non-synonymous polymorphisms. Here, we describe a program designed to quickly assimilate all available data for SNPs or locations entered by the user, called ASSIMILATOR. Importantly, the ability to enter SNPs using base pair position will allow the interrogation of novel variants identified, for example, by the 1000 Genomes project (http://www.1000genomes.org) even if an rs number has not yet been assigned. We also validate the method by interrogating SNPs in two regions: one associated with colorectal cancer (Pomerantz *et al.*, 2009) and one with type II diabetes (T2D) (Gaulton *et al.*, 2010). We show that, based on the information drawn together by ASSIMILATOR, we would have prioritized the subsequently confirmed causal SNPs for functional investigation from both previous studies.

2 METHODS

Written in Perl, ASSIMILATOR retrieves, queries and processes information for the desired SNPs from the UCSC Genome Browser's public MySQL database and displays this in a simplified, user-friendly manner. All available ENCODE tracks are queried in addition to predefined tracks, such as mRNAs, ESTs and CpG islands. In addition, eQTL data hosted by the Pritchard laboratories (http://eqtl.uchicago.edu), PolyPhen2 functional annotation (Adzhubei *et al.*, 2010) and SNP location relative to the gene are displayed. Multiple systems have been designed to improve the efficiency of data retrieval such as an XML-based track database, which minimizes the number of database queries and multi-threading support to query multiple SNPs simultaneously, reducing processing time with minimal reduction in individual performance.

The output can be viewed in a standard web browser and allows the user to quickly identify SNPs, which could be functionally important. To add extra functionality, the ability to view selected SNPs in NCBIs dbSNP (Sherry *et al.*, 2001) and in the UCSC Genome Browser has been incorporated into the output. To efficiently display features for a SNP in the UCSC Genome Browser, only tracks that contain features in the SNP region are displayed. The user interface has been designed to allow further mining of the output (Fig. 1) to display information from the multiple cell types and links to external data. This includes the ability to view the detailed experimental data thereby allowing users to assess the biological relevance of the results in the

*To whom correspondence should be addressed.



Location Information			Standard					Expression					Regulation						
SNP ID	SNP Position	Information	Conservation	Human ESTs	Human mRNAs	Gencode Genes	Relative Location	Affy RNA Loc	Caltech RNA-seq	CSHL Sm RNA-seq	GIS PET Loc	RIKEN CAGE Loc	Broad Histone	Open Chromatin	HAIB TFBS	UW Histone	UW DNase DGF	UW DNaseI HS	Yale TFBS
rs6983267	chr8:128482487-128482487	snp130	Yes	Yes	Yes			Yes	Yes		Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
rs10808556	chr8:128482329-128482329	snp130		Yes	Yes				Yes		Yes		Yes	Yes	Yes	Yes	Yes	Yes	Yes
rs4871788	chr8:128490967-128490967	snp130		Yes	Yes				Yes		Yes	Yes	Yes	Yes		Yes		Yes	
rs3847137	chr8:128483680-128483680	snp130											Yes	Yes		Yes		Yes	
rs2060776	chr8:128489299-128489299	snp130											Yes						
rs4276648	chr8:128496554-128496554	snp130											Yes	Yes		Yes			
rs4871022	chr8:128496902-128496902	snp130											Yes	Yes		Yes		Yes	
rs10956369	chr8:128492999-128492999	snp130											Yes	Yes		Yes		Yes	
rs7837644	chr8:128492580-128492580	snp130											Yes					Yes	
rs871135	chr8:128495575-128495575	snp130											Yes						
rs10505477	chr8:128476625-128476625	snp130											Yes					Yes	
rs10505474	chr8:128486686-128486686	snp130											Yes						
rs7837328	chr8:128492309-128492309	snp130											Yes			Yes			
rs10956368	chr8:128492832-128492832	snp130										s	Yes	Yes					
rs7837626	chr8:128492523-128492523	snp130		Yes	Yes				Yes	Yes	Yes		Yes	Yes				Yes	

wgEncodeRikenCage (rs6983267)

1. Track: ENCODE RIKEN CAGE Tags (PolyA- RNA in prostate whole cell)
Date Unrestricted: 2009-09-09
Name: 00003022001131200022003113
Position: chr8:128482488-128482513
Strand: -
Score: 200

2. Track: ENCODE RIKEN CAGE Tags (PolyA- RNA in prostate whole cell)
Date Unrestricted: 2009-09-09
Name: 00003022001131200022003113
Position: chr8:128482488-128482513
Strand: -
Score: 200

Location Information			Standard					Expression							Regulation								VarRep
SNP ID	SNP Position	Information	Conservation	Human ESTs	Human mRNAs	Gencode Genes	Relative Location	Affy RNA Loc	Caltech RNA-seq	GIS PET Loc	GIS RNA-seq	HudsonAlpha RNA-seq	RIKEN CAGE Loc	UW Affy Exon	Broad Histone	Open Chromatin	HAIB TFBS	SUNY RBP	UW Histone	UW DNase DGF	UW DNaseI HS	Yale TFBS	Common Cell CNV
	chr8:128482487		Yes	Yes	Yes			Yes	Yes	Yes			Yes		Yes	Yes	Yes		Yes	Yes	Yes	Yes	
rs7903146	chr10:114748339-114748339	snp130		Yes	Yes	Yes	Intronic		Yes	Yes	Yes	Yes		Yes	Yes	Yes		Yes			Yes		Yes

Fig. 1. Examples of ASSIMILATOR output showing results for (**a**) Pomerantz *et al*. with the causal SNP highlighted and (**b**) Gaulton *et al*. showing the evidence that the SNP is in a region of open chromatin. In addition, an example of results for a SNP without an rs number, as might be the case for novel SNPs identified via the 1000 Genomes project (http://www.1000genomes.org), is shown.

context of the thresholds and criteria used. ASSIMILATOR automatically queries any new tracks appearing from the ENCODE project on UCSC and includes these in the analysis. To further ensure ASSIMILATOR stays up to date, an option is available, which searches all UCSC database versions for ENCODE tracks and automatically uses the latest suitable version [currently March. 2006 (NCBI36/hg18)]. The ENCODE data release policy places restrictions on the publication of ENCODE data; therefore, the date at which the data becomes unrestricted is also displayed to aid the user.

To analyse the data, a hierarchical approach can be employed by the user, where isolated evidence for conservation across species, evidence of histone modification or mapping to a methylated region might be assigned a low weighting by the user; conversely, consistent evidence for a region being active, such as evidence for histone modification, DNase-1 hypersensitivity and open chromatin in the same cell line, coupled with evidence that a SNP lies within a transcription factor binding site (TFBS) would receive a higher weighting and could help to prioritize that SNP for functional work and may inform the design of such studies.

3 RESULTS

To verify the usefulness of ASSIMILATOR, we used information from a published study by Pomerantz *et al.* who found that an intergenic SNP, rs6983267, associated with colorectal cancer, showed functional evidence for interaction with the *MYC* gene (Pomerantz *et al.*, 2009). We used the SNP Annotation and Proxy Search (SNAP) tool (Johnson *et al.*, 2008) to generate a list of SNPs highly correlated with rs6983267 ($r^2 > 0.8$). This generated a list of 15 SNPs that were subsequently used as the input to ASSIMILATOR. The results are shown in Figure 1a clearly indicating that rs6983267 has the strongest a priori evidence of function. Not only is it in an active region of the genome, but also it is one of only two SNPs to lie in a TFBS. Additionally, ASSIMILATOR correctly identified the same TFBS as the published data.

Similarly, a recent study by Gaulton *et al.* (2010) looking at open chromatin across the genome identified a SNP associated with T2D in an open region. As a further proof of concept, supplying ASSIMILATOR with the same SNP revealed three lines of evidence showing bioinformatically that the SNP was in a region of open chromatin (Fig. 1b). This selection was achieved quickly and easily using our programme.

4 CONCLUSIONS

ASSIMILATOR provides a user-friendly interface with which to collate and assess the wealth of experimental evidence available for SNPs in order to prioritize efficiently for functional studies. ASSIMILATOR does not try to make assumptions about the likelihood of a SNP being functional and as such allows the user to make their own judgements about the candidacy of a SNP. ASSIMILATOR will also quickly and easily incorporate new data added to the ENCODE project ensuring that it maintains its relevance. With the wealth of information emerging from genome annotation studies, the task of manually mining the thousands of data points would be daunting. Here, we provide a one-stop solution that quickly and efficiently allows the user to view only relevant studies for their SNPs of interest and to mine that data with ease.

We have validated the program using published data and have shown that it allows the correct prioritization of a SNP subsequently shown to be the causal variant in a region associated with colorectal cancer. It thus provides an efficient portal to gather the essential information on which to base decisions regarding priorities for functional work. We have made ASSIMILATOR freely available through our web site as a download and we are also developing a web-based interface which will be found at the same location.

Funding: Arthritis Research UK (grant no. 17752).

Conflict of Interest: none declared.

REFERENCES

Adzhubei,I.A. *et al*. (2010) A method and server for predicting damaging missense mutations. *Nat. Methods*, **7**, 248–249.

Barton,A. *et al*. (2008) Rheumatoid arthritis susceptibility loci at chromosomes 10p15, 12q13 and 22q13. *Nat. Genet.*, **40**, 1156–1159.

Birney,E. *et al*. (2007) Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature*, **447**, 799–816.

ENCODE Project Consortium (2004) The ENCODE (ENCyclopedia Of DNA Elements) Project. *Science*, **306**, 636–640.

Gaulton,K.J. *et al*. (2010) A map of open chromatin in human pancreatic islets. *Nat. Genet.*, **42**, 255–259.

Johnson,A.D. *et al*. (2008) SNAP: a web-based tool for identification and annotation of proxy SNPs using HapMap. *Bioinformatics*, **24**, 2938–2939.

Kent,W.J. *et al*. (2002) The human genome browser at UCSC. *Genome Res*., **12**, 996–1006.

Pomerantz,M.M. *et al*. (2009) The 8q24 cancer risk variant rs6983267 shows long-range interaction with MYC in colorectal cancer. *Nat. Genet.*, **41**, 882–884.

Sherry,S.T. *et al*. (2001) dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res*., **29**, 308–311.

Thomson,W. *et al*. (2007) Rheumatoid arthritis association at 6q23. *Nat. Genet.*, **39**, 1431–1433.

Wellcome Trust Case Control Consortium (2007) Genome-wide association study of 14 000 cases of seven common diseases and 3,000 shared controls. *Nature*, **447**, 661–678.

Publication 2: Novel rheumatoid arthritis susceptibility locus at 22q12 identified in an extended UK genome-wide association study

ARTHRITIS & RHEUMATOLOGY
Vol. 66, No. 1, January 2014, pp 24–30
DOI 10.1002/art.38196

Novel Rheumatoid Arthritis Susceptibility Locus at 22q12 Identified in an Extended UK Genome-Wide Association Study

Gisela Orozco,[1] Sebastien Viatte,[1] John Bowes,[1] Paul Martin,[1] Anthony G. Wilson,[2] Ann W. Morgan,[3] Sophia Steer,[4] Paul Wordsworth,[5] Lynne J. Hocking,[6] UK Rheumatoid Arthritis Genetics Consortium, Wellcome Trust Case Control Consortium, Biologics in Rheumatoid Arthritis Genetics and Genomics Study Syndicate Consortium, Anne Barton,[1] Jane Worthington,[1] and Stephen Eyre[1]

Objective. **The number of confirmed rheumatoid arthritis (RA) loci currently stands at 32, but many lines of evidence indicate that expansion of existing genome-wide association studies (GWAS) enhances the power to detect additional loci. This study was undertaken to extend our previous RA GWAS in a UK cohort, adding more independent RA cases and healthy controls, with the aim of detecting novel association signals for susceptibility to RA in a homogeneous UK cohort.**

Methods. **A total of 3,223 UK RA cases and 5,272 UK controls were available for association analyses, with the extension adding 1,361 cases and 2,334 controls to the original GWAS data set. The genotype data for all RA cases were imputed using the Impute program version 2. After stringent quality control thresholds were applied, 3,034 cases and 5,271 controls (1,831,729 single-nucleotide polymorphisms [SNPs]) were available for analysis. Association testing was performed using Plink software.**

Results. **The analyses indicated a suggestive association with susceptibility to RA ($P < 0.0001$) for 6 novel RA loci that have been previously found to be associated with other autoimmune diseases; these 6 SNPs were genotyped in independent samples. Two of the associated loci were validated, one of which was associated with RA at genome-wide levels of significance in the combined analysis, identifying a novel RA locus at 22q12 ($P = 6.9 \times 10^{-9}$). In addition, most of the previously known RA susceptibility loci were confirmed to be associated with RA, and for 16 of the loci, the strength of the association was increased.**

Conclusion. **This study identified a new RA locus mapping to 22q12. These results support the notion that increasing the power of GWAS enhances novel gene discovery.**

Understanding the genetic component of susceptibility to rheumatoid arthritis (RA) will increase our

Funding to generate the genome-wide association study data was provided by Arthritis Research UK (grant 17552), Sanofi-Aventis, and the European Union Seventh Framework Programme (project POCEMON; FP7-ICT-2007-216088). Genotyping was performed at the University of Milan. Dr. Orozco's work was supported by the Wellcome Trust (Research Career Development Fellowship 095684/Z/11/A). Dr. Viatte's work was supported by the Swiss Foundation for Medical-Biological Scholarships, which is managed by the Swiss National Science Foundation (SSMBS grant PASMP3_134380, funded by a donation from Novartis to the Foundation).

[1]Gisela Orozco, PhD, Sebastien Viatte, MD, PhD, John Bowes, PhD, Paul Martin, MPhil, Anne Barton, PhD, FRCP, Jane Worthington, PhD, Stephen Eyre, PhD: Arthritis Research UK Epidemiology Unit, Centre for Musculoskeletal Research, Institute of Inflammation and Repair, University of Manchester, and NIHR Manchester Musculoskeletal Biomedical Research Unit, Manchester Academic Health Science Centre, Manchester, UK; [2]Anthony G. Wilson, MB, PhD, FRCP: University of Sheffield, Sheffield, UK; [3]Ann W. Morgan, PhD, FRCP: University of Leeds, Leeds, UK; [4]Sophia Steer, PhD, MRCP: Kings College Hospital National Health Service Foundation Trust, London, UK; [5]Paul Wordsworth, MA, MB, BS, FRCP: NIHR Oxford Musculoskeletal Research Unit and Botnar Research Centre, Nuffield Orthopaedic Centre, Oxford, UK; [6]Lynne J. Hocking, PhD: University of Aberdeen, Aberdeen, UK.

Dr. Martin has received consulting fees, speaking fees, and/or honoraria from Eli Lilly, Pfizer, AbbVie, and Merck (less than $10,000 each).

Address correspondence to Stephen Eyre, PhD, University of Manchester, Arthritis Research UK Epidemiology Unit, Centre for Musculoskeletal Research, Institute of Inflammation and Repair, Manchester Academic Health Science Centre, Stopford Building, Oxford Road, Manchester M13 9PT, UK. E-mail: stephen.eyre@manchester.ac.uk.

Submitted for publication August 22, 2012; accepted in revised form September 10, 2013.

Table 1. Previously confirmed rheumatoid arthritis (RA) loci association results in the original WTCCC association analysis and the expanded UK RA GWAS*

			WTCCC study			Expanded UK RA GWAS		
Chr.	SNP	Gene	Proxy	*P*	OR (95% CI)	Proxy	*P*	OR (95% CI)
1	rs3890745	*TNFRSF14*		8.47×10^{-6}	0.82 (0.75–0.89)		1.43×10^{-6}	0.85 (0.79–0.91)
1	rs2476601	*PTPN22*	rs6679677	2.60×10^{-25}	1.90 (1.68–2.15)		4.87×10^{-33}	1.77 (1.61–1.95)
1	rs11586238	*CD2, CD58*		4.13×10^{-4}	1.19 (1.08–1.31)	rs4271251	6.47×10^{-3}	1.11 (1.03–1.19)
1	rs12746613	*FCGR2A*		0.04	1.14 (1.01–1.29)		0.03	1.11 (1.01–1.22)
1	rs10919563	*PTPRC*		0.003	0.82 (0.72–0.94)		7.18×10^{-5}	0.82 (0.74–0.9)
2	rs13031237	*REL*	None			rs13031721	0.29	1.04 (0.97–1.11)
2	rs934734	*SPRED2*		0.10	0.93 (0.86–1.01)		0.11	0.95 (0.89–1.01)
2	rs10865035	*AFF3*	rs9653442	5.48×10^{-4}	1.16 (1.07–1.26)	rs1160542	1.37×10^{-5}	1.15 (1.08–1.23)
2	rs7574865	*STAT4*	rs11893432	0.02	1.13 (1.02–1.25)	rs10181656	6.64×10^{-4}	1.14 (1.06–1.23)
2	rs1980422	*CD28*		4.80×10^{-3}	1.15 (1.04–1.26)		1.75×10^{-4}	1.15 (1.07–1.24)
2	rs3087243	*CTLA4*		0.09	0.93 (0.86–1.01)		0.02	0.92 (0.87–0.99)
3	rs13315591	*PXK*		0.20	1.10 (0.95–1.26)	None		
4	rs874040	*RBPJ*	None			rs6448432	3.88×10^{-7}	1.19 (1.11–1.28)
4	rs6822844	*IL2, IL21*	None			rs62322744	6.42×10^{-3}	1.18 (1.05–1.32)
5	rs6859219	*ANKRD55, IL6ST*		5.5×10^{-6}	0.78 (0.70–0.87)	None		
5	rs26232	*C5orf30*	rs556560	2.46×10^{-4}	0.85 (0.78–0.93)	rs556560	9.64×10^{-5}	0.88 (0.82–0.94)
6	rs6910071	*HLA–DRB1*	rs6457617	3.49×10^{-79}	0.44 (0.40–0.48)	rs3763309	1.50×10^{-124}	2.3 (2.14–2.46)
6	rs548234	*PRDM1*		0.01	1.12 (1.02–1.22)		1.24×10^{-3}	1.12 (1.04–1.19)
6	rs6920220	*TNFAIP3*		6.11×10^{-6}	1.25 (1.13–1.37)		3.11×10^{-8}	1.23 (1.14–1.32)
6	rs394581	*TAGAP*		5.86×10^{-3}	0.88 (0.80–0.96)	None		
6	rs3093023	*CCR6*	rs6907666	0.05	1.09 (1–1.18)	rs3093024	1.88×10^{-3}	1.11 (1.04–1.18)
7	rs10488631	*IRF5*	rs12531711	0.03	1.16 (1.02–1.31)		3.10×10^{-3}	1.16 (1.05–1.28)
8	rs2736340	*BLK*		8.01×10^{-3}	1.14 (1.03–1.25)		0.05	1.07 (1–1.15)
9	rs2812378	*CCL21*		1.14×10^{-3}	1.15 (1.06-1.26)	None		
9	rs3761847	*TRAF1, C5*		0.80	0.99 (0.91–1.07)		0.19	1.04 (0.98–1.11)
10	rs2104286	*IL2RA*		7.06×10^{-6}	0.81 (0.73–0.89)		1.46×10^{-6}	0.84 (0.78–0.9)
10	rs4750316	*PRKCQ*		5.22×10^{-5}	0.80 (0.72–0.89)		1.55×10^{-4}	0.85 (0.78–0.93)
11	rs540386	*TRAF6*		0.04	0.88 (0.77–0.99)	rs1046864	0.01	0.89 (0.8–0.98)
12	rs1678542	*KIF5A*		2.81×10^{-5}	0.83 (0.76–0.91)		9.75×10^{-8}	0.83 (0.78–0.89)
20	rs4810485	*CD40*		0.07	0.91 (0.83–1.01)	rs1569723	0.22	0.95 (0.89–1.03)
22	rs3218253	*IL2RB*		1.88×10^{-4}	1.19 (1.09–1.31)		2.51×10^{-4}	1.15 (1.07–1.23)

* WTCCC = Wellcome Trust Case Control Consortium; GWAS = genome-wide association study; Chr. = chromosome; SNP = single-nucleotide polymorphism; OR = odds ratio; 95% CI = 95% confidence interval.

knowledge of the disease process and has the potential to inform new approaches to disease management. For example, the identification of disease-associated genetic variations, which are presumed to cause modified immune responses and precede the onset of disease symptoms, could inform stratification of patients into more phenotypically homogeneous subgroups and provide testable hypotheses regarding response to treatment. The use of genome-wide association studies (GWAS) has been remarkably successful in locating novel genetic loci associated with RA, and there are now more than 30 gene regions that have been confirmed as RA susceptibility loci, but, in total, they account for fewer than 50% of the total genetic heritability (1,2). It is likely that there are more common variants of small effect size that could be identified by increasing the power of the GWAS through the use of larger sample sizes. Indeed, this approach has been used successfully in a number of autoimmune diseases, such as type 1 diabetes (3) and inflammatory bowel disease (4,5), resulting in a more complete picture of the genetic background.

In the present study, we used the data set from an RA GWAS in the UK (6) and increased the sample sizes of the RA cases and healthy controls by 75% and 80%, respectively. With this extended data set, together with the data obtained in a validation study of the UK cohort, we were able to discover potential novel RA risk loci in the UK population. This study constitutes the largest UK-only GWAS to date, and the results will enhance the power to investigate whether population heterogeneity exists, i.e., whether different genes are associated with RA in different populations.

PATIENTS AND METHODS

Genotype data were available for 1,862 RA cases and 2,938 controls from the original Wellcome Trust Case Control

Consortium (WTCCC) study (6); these genotypes were obtained using the Affymetrix 500K array. Together with this data set, we added GWAS genotype data for a further 2,334 UK controls and 1,361 UK RA cases (sample call rate >95%). The additional samples were from 5 different studies, with genotypes obtained using a range of GWAS arrays. Members of the WTCCC, UK Rheumatoid Arthritis Genetics Group, and Biologics in Rheumatoid Arthritis Genetics and Genomics Study Syndicate Consortiums and details on the platforms used are given in the Supplementary Materials and Supplementary Table 1, available on the *Arthritis & Rheumatology* web site at http://onlinelibrary.wiley.com/doi/10.1002/art.38196/abstract.

Since each GWAS array contains different single-nucleotide polymorphisms (SNPs), the first stage of the current study was to impute genotypes to generate data for a common set of SNPs, thus allowing combined analysis of the data from the 5 different studies. The genotypes of the patients in each RA case cohort were imputed using the Impute program (version 2) (7), based on 2 reference panels, the 1000 Genomes Project pilot data and HapMap3. The genotypes of control subjects were imputed using the same 1000 Genomes Project reference panel, with imputation performed using MaCH software (8). Stringent quality control (QC) thresholds were applied to individual cohorts and then to the merged data set (see Supplementary Figure 1, available on the *Arthritis & Rheumatology* web site at http://onlinelibrary.wiley.com/doi/10.1002/art.38196/abstract). Self-reported information on each subject's ethnicity was used to exclude non-Caucasians from the analyses.

Identity-by-descent analysis was performed in Plink statistical software. Pairs of samples with genome identity (pi-hat values) greater than 0.9 were reported to be duplicate samples, and those with pi-hat values greater than 0.2 were considered to be closely related. Case and control genotype data were compared using logistic regression, performed in Plink (9). Novel SNPs not previously found to be associated with RA in the UK population were selected for validation if the association *P* value was less than 0.0001, and if the region had been previously found to be associated with another autoimmune disease. The SNPs satisfying these criteria were genotyped in independent UK samples (4,726 cases and 2,625 controls), using Sequenom technology. Additional control data were available using non–autoimmune disease cases (n = 7,670) from the WTCCC study (6).

A regional association plot for the 22q12 locus was created using LocusZoom (http://csg.sph.umich.edu/locuszoom/). Bioinformatics analysis was performed using publicly available functional annotation files (http://genome.ucsc.edu/) and our own custom query SQL-code (Assimilator; available at http://assimilator.mhs.manchester.ac.uk/cgi-bin/assimilator.pl) (10).

RESULTS

Forty-two duplicate samples were removed after identity-by-descent analysis. There were 14 pairs of individual samples with pi-hat values greater than 0.2, and thus 1 sample from each pair was removed due to possible relatedness. A total of 1,831,729 SNPs passed QC and were included in the association analysis. In the expanded GWAS, 3,034 RA cases and 5,271 controls passed QC.

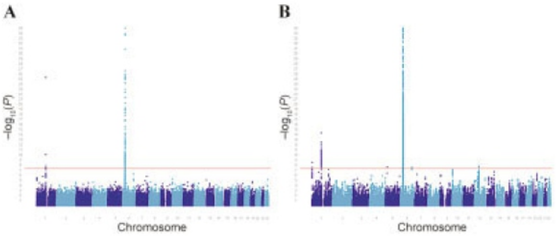


Figure 1. Manhattan plots showing *P* values for genetic association with rheumatoid arthritis (RA) in the original Wellcome Trust Case Control Consortium genome-wide association study (GWAS) **(a)** and the expanded UK GWAS **(b)**. Panels are truncated at a $-\log_{10} P$ value of 35. Single-nucleotide polymorphisms having an association with RA at genome-wide levels of significance are shown above the horizontal red line.

The genetic inflation factor lambda (λ_{GC}) was 1.06, implying that there was a low possibility of false-positive associations attributable to population stratification, genotyping errors, or other artifacts. For most of the previously known RA susceptibility loci (2), we confirmed their association with RA in the present study (Table 1). Interestingly, when compared to the original WTCCC study (6), the extended RA GWAS showed that the strength of the association with RA susceptibility was increased for 16 of the loci in the UK population (*TNFAIP3, STAT4*, *PTPN22, HLA–DRB1, PTPRC, IL2/IL21, C5orf30, CD247*, *CTLA4, RBPJ, PRDM1, CCR6, IRF5, TRAF6, KIF5A*, and *CD28*). For example, the association of *TNFAIP3* with RA risk reached genome-wide levels of significance in this expansion of the WTCCC analysis, and the robustly validated locus *STAT4*, for which no significant association had been found in the original WTCCC study, was found to be associated with RA in the present study (Figure 1).

Similar to the findings in the original WTCCC study, some of the known RA loci (*CD40, REL, SPRED2, BLK*, and *TRAF1*/*C5*) were not found to have an association with RA in the present study. Finally, no proxies were available for 4 of the known RA loci.

Six of the SNPs with suggestive evidence of association with RA in the present study and with additional evidence of association in other studies were selected for validation analysis (Table 2). Evidence of an association with susceptibility to RA was found for rs1043099 (on chromosome 22q12) in this validation study ($P = 2.7 \times 10^{-3}$), and the association exceeded

Table 2. Candidate rheumatoid arthritis (RA) susceptibility loci previously associated with other autoimmune diseases and showing suggestive evidence of association with RA in the expanded UK RA GWAS*

				Expanded UK RA GWAS					Validation study				Combined analysis	
				MAF					MAF					
Chr.	SNP	Locus	Assoicated autoimmune disease	RA	Controls	P†‡	OR (95% CI)‡	Proxy/ comment	RA	Controls	P	OR (95% CI)	P	OR (95% CI)
7	rs1047022	*SNX10*	T1D, IBD	0.15	0.19	6.4×10^{-7}	0.80 (0.74–0.88)	Failed QC						
11	rs6421571	*CXCR5*	PBC	0.17	0.19	6.3×10^{-5}	0.84 (0.78–0.92)		0.18	0.19	0.02	0.90 (0.82–0.98)	1.8×10^{-5}	0.88 (0.83–0.93)
12	rs11181399	*YAF2*	MS	0.40	0.45	2.5×10^{-7}	0.8 (0.79–0.90)	rs7954523	0.39	0.39	0.96	0.99 (0.93–1.08)	7.01×10^{-5}	0.91 (0.87–0.95)
16	rs7187962	*CNTAP4*	T1D	0.27	0.31	8.6×10^{-6}	0.85 (0.79–0.91)	rs6564350§	0.30	0.31	0.40	0.97 (0.89–1.05)		
19	rs6417247	*CDC37*	IBD	0.30	0.33	9.1×10^{-5}	0.87 (0.82–0.93)	rs8112449	0.32	0.33	0.20	0.95 (0.88–1.03)	1.5×10^{-3}	0.93 (0.88–0.97)
22	rs1043099	*GATSL3*	CD, T1D	0.18	0.21	7.8×10^{-6}	0.83 (0.77–0.90)		0.19	0.21	2.70×10^{-3}	0.87 (0.80–0.95)	6.9×10^{-9}	0.84 (0.79–0.89)

* Associations were identified at the significance level of $P < 0.0001$. Chr. = chromosome; MAF = minor allele frequency; T1D = type 1 diabetes; IBD = inflammatory bowel disease; QC = quality control; PBC = primary biliary cirrhosis; MS = multiple sclerosis; CD = celiac disease.

† Corrected P values (corrected for genetic inflation factor lambda) were as follows: for rs1047022, $P = 1.42 \times 10^{-6}$; for rs6421571, $P = 1.07 \times 10^{-4}$; for rs11181399, $P = 5.86 \times 10^{-7}$; for rs7187962, $P = 1.628 \times 10^{-5}$; for rs6417247, $P = 1.50 \times 10^{-4}$; for rs1043099, $P = 1.49 \times 10^{-5}$.

‡ Association results in the original Wellcome Trust Case Control Consortium study were as follows: for rs1047022, $P = 1.106 \times 10^{-6}$, odds ratio (OR) 0.74 (95% confidence interval [95% CI] 0.66–0.84); for rs6421571, $P = 0.02$, OR 0.88 (95% CI 0.79–0.98); for rs11181399, $P = 3.98 \times 10^{-5}$, OR 0.83 (95% CI 0.76–0.91); for rs7187962, $P = 7.07 \times 10^{-6}$, OR 0.80 (95% CI 0.73–0.88); for rs6417247, $P = 2.31 \times 10^{-3}$, OR 0.87 (95% CI 0.79–0.95); for rs1043099, $P = 1.65 \times 10^{-4}$, OR 0.81 (95% CI 0.72–0.90).

§ Single-nucleotide polymorphism (SNP) rs6564350 was not present in the expanded UK RA genome-wide association study (GWAS) data set, and therefore combined analysis of this SNP could not be performed.

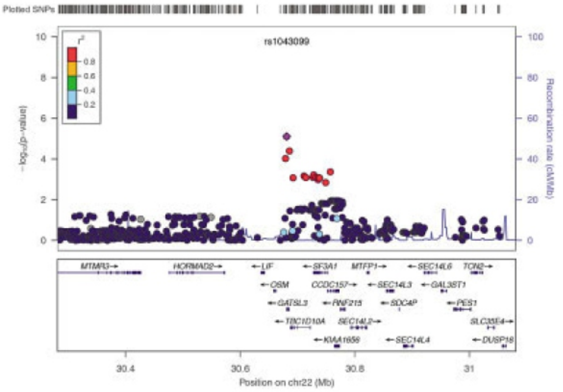


Figure 2. Regional plot of association with rheumatoid arthritis (RA) at chromosome 22q12. The *P* values for association ($-\log_{10}$ values) of each single-nucleotide polymorphism (SNP) are plotted against their physical position on chromosome 22 (top panel). Estimated recombination rates from the 1000 Genomes Project population show the local linkage disequilibrium (LD) structure (middle panel). Different colors indicate the LD of each SNP with rs1043099, based on pairwise r^2 values from the 1000 Genomes Project. Gene annotations are shown in the lower panel.

genome-wide significance thresholds in the combined analysis ($P = 6.9 \times 10^{-9}$) (Figure 2). A SNP on chromosome 11, near *CXCR5* and in strong linkage disequilibrium (LD) ($r^2 > 0.9$) with a SNP previously associated with RA at genome-wide significance levels (11), was also associated with RA in the validation cohort ($P = 0.02$), and there was suggestive evidence of association in the combined analysis ($P = 1.8 \times 10^{-5}$).

In further analyses, we interrogated publicly available functional annotation data (10,12) and found evidence to indicate that rs1043099 and its correlated SNPs ($r^2 > 0.8$) may have regulatory activity in RA. First, the SNP alleles were found to be associated with expression levels of the *SF3A1* gene (which encodes subunit 1 of the splicing factor 3a protein complex) (http://eqtl.uchicago.edu/cgi-bin/gbrowse/eqtl/) (Mangravite LM, et al: unpublished observations). Second, several of the SNPs map to sites of transcription factor binding, histone modification, and open chromatin (Table 3), suggesting that these SNPs could influence gene transcription.

We then stratified the expanded UK GWAS and validation data sets according to the presence or absence of anti–cyclic citrullinated peptide (anti-CCP) antibodies in RA patients. The proportion of anti-CCP–positive RA patients in the original WTCCC study, expanded UK GWAS, and validation study was 79%, 73.5%, and 67%, respectively. SNP rs1043099 was significantly associated with both anti-CCP–positive RA and anti-CCP–negative RA. In contrast, SNP rs6421571 was associated with anti-CCP–positive RA only (see Supplementary Table 2, available on the *Arthritis & Rheumatology* web site at http://onlinelibrary.wiley.com/doi/10.1002/art.38196/abstract). Similar results were obtained when patients were stratified by the presence of rheumatoid

Table 3. Potential regulatory role of the rheumatoid arthritis–associated single-nucleotide polymorphism (SNP) rs1043099 and its proxies*

SNP	Relative location	Gene	Transcribed region	Histone modifications	TFBS	DNase I HS	FAIRE	CTCF binding	eQTL
rs2108093	3′	*GATSL3*	Yes	Yes	Yes	Yes			Yes
rs1043099	Exonic, 3′-UTR	*GATSL3*	Yes	Yes		Yes		Yes	Yes
rs4823085	5′	*GATSL3*	Yes	Yes	Yes	Yes	Yes	Yes	Yes
rs929454	Intronic	*TBC1D10A*	Yes	Yes		Yes		Yes	Yes
rs4820831	Intronic	*TBC1D10A*	Yes	Yes	Yes	Yes		Yes	Yes
rs740219	Intronic	*TBC1D10A*	Yes	Yes		Yes		Yes	Yes
rs5753071	3′	*SF3A1*	Yes	Yes	Yes	Yes			Yes
rs10376	Exonic, 3′-UTR	*SF3A1*	Yes	Yes	Yes	Yes			Yes
rs2041199	Intronic	*SF3A1*	Yes	Yes	Yes				Yes
rs4339043	Intronic	*SF3A1*	Yes	Yes	Yes	Yes			Yes
rs5749066	Intronic	*SF3A1*	Yes	Yes	Yes	Yes			Yes
rs5753080	Intronic	*SF3A1*	Yes	Yes					Yes
rs10427610	Intronic	*SF3A1*	Yes	Yes	Yes	Yes	Yes		Yes
rs4820008	Intronic	*SF3A1*	Yes	Yes					Yes
rs737950	Intronic	*CCDC157*	Yes	Yes					Yes
rs5749078	Intronic	*CCDC157*	Yes	Yes		Yes			Yes
rs9619104	3′	*CCDC157*	Yes	Yes		Yes			Yes

* Proxies of rs1043099 were correlated at $r^2 > 0.8$. Results are the summary output from the Assimilator bioinformatics analysis. TFBS = transcription factor binding site; HS = hypersensitive sites; FAIRE = open chromatin by formaldehyde-assisted isolation of regulatory elements; CTCF = CCCTC binding factor; eQTL = expression quantitative trait loci; 3′-UTR = 3′-untranslated region.

factor (see Supplementary Table 3, available on the *Arthritis & Rheumatology* web site at http://onlinelibrary.wiley.com/doi/10.1002/art.38196/abstract).

DISCUSSION

In this study, we discovered a new RA risk locus on chromosome 22q12, rs1043099, for which the association with susceptibility to RA reached genome-wide levels of significance (combined $P = 6.9 \times 10^{-9}$). This locus has previously been associated with other autoimmune diseases, including type 1 diabetes (3) and inflammatory bowel disease (5,13). The SNP lies within a gene of unknown function, *GATSL3*. It will be interesting to better fine-map this region in RA, as well as in other autoimmune diseases, to help determine which gene is causal. Using bioinformatics analysis, we showed that rs1043099 and its correlated SNPs have potential regulatory activity. Interestingly, evidence suggests that these SNPs act as expression quantitative trait loci for the *SF3A1* gene. *SFA1* is involved in messenger RNA splicing, but the particular role that this gene might have in the pathogenesis of RA has not yet been explored. Further functional studies will be required to determine which of the SNPs is causal and to elucidate the mechanisms of action of each SNP.

Of the remaining SNPs tested in the validation samples, evidence of increased strength of the association with susceptibility to RA was found only for rs6421571 on chromosome 11q23 (combined $P = 1.8 \times 10^{-5}$). This SNP maps 5′ to the *CXCR5* gene, which is a chemokine involved in B cell migration and localization, and has previously been associated with primary biliary cirrhosis (14). Different SNPs in the same region have been found to be associated with multiple sclerosis (15) and RA (11). The SNP identified as being significantly associated with RA and celiac disease in a combined analysis, rs10892279, is adjacent to the gene *DDX6*, and although it is >130 kb from rs6421571, the 2 SNPs are strongly correlated ($r^2 > 0.9$). Therefore, which of these 2 strong candidate genes will ultimately be found to be causal requires further fine-mapping and functional studies.

The added power provided by this study strengthens the evidence for an association with RA for 16 loci that were already confirmed to be associated with RA in the UK population (2). The most striking findings supporting an association with RA within the UK population, as evidenced by an increase in the significance of the association from the original WTCCC study to the expanded GWAS, were seen for the loci at *PTPRC* (increasing from $P = 0.003$ to $P = 7 \times 10^{-5}$), *TNFAIP3* (increasing from $P = 6 \times 10^{-6}$ to $P = 3 \times 10^{-8}$), and *KIF5A* (increasing from $P = 2 \times 10^{-5}$ to $P = 9 \times 10^{-8}$). *RBPJ*, not available for analysis in the original WTCCC study, showed evidence of a strong association with RA in this expanded GWAS ($P = 3 \times 10^{-7}$).

There was little evidence of an association with other previously confirmed RA loci in this expanded UK data set, including *FCGR2A, REL, SPRED2, CTLA4, BLK, TRAF1*, and *CD40*. This could be attributed to a number of factors, but the most likely explanation is that the current study was underpowered to detect such associations. Although we performed a relatively large discovery study, the power to detect all of the associations was limited (average power across SNPs with a minor allele frequency >5% was 47% for an odds ratio [OR] of 1.1, and >90% for an OR of >1.2), and therefore failure to detect a signal could simply be the result of stochastic variations leading to a false-negative result. Indeed, it is well established that even the most strongly associated signals across multiple diseases are not found to be consistently associated in all cohorts, and often an accumulation of evidence is required in many thousands of samples. In this regard, a recent meta-analysis failed to detect the association between the 22q12 locus and RA that we identified in the current study (for rs929454, $P = 0.63$, $r^2 = 0.91$ for LD with rs1043099) (2).

Interestingly, most of the confirmed RA loci for which we did not find any association with RA in the expanded UK data set (*FCGR2A, REL, SPRED2, CTLA4, BLK, TRAF1*, and *CD40*) were found to have an association with anti-CCP–positive RA, but not anti-CCP–negative RA, in a recent study, which was the largest anti-CCP–stratified RA sample size studied to date (16). We found very modest evidence of association with anti-CCP–positive RA for only 2 of the above-mentioned variants ($P = 0.03$ for *FCGR2A*, and $P = 0.03$ for *BLK*), and therefore it is likely that the present study was underpowered to detect serotype-specific effects at these RA loci.

This study has added to the argument that increasing the size, and consequently the power, of RA GWAS can prove fruitful in the discovery of novel genes. Increasing the number of known disease loci will facilitate the estimation of disease risk, potentially allowing early intervention in high-risk groups, possibly informing prognosis, and, ultimately, aiding in the discovery of novel targets for pharmacologic therapy. Future work will involve adding this new extended UK data to existing meta-analyses of RA case data with the aim of expanding our knowledge of the common genetic variants predisposing to RA.

ACKNOWLEDGMENTS

We would like to thank Edward Flynn for preparing and genotyping the UK Rheumatoid Arthritis Group samples. We thank Eleftheria Zeggini for providing the imputed data for healthy controls.

AUTHOR CONTRIBUTIONS

All authors were involved in drafting the article or revising it critically for important intellectual content, and all authors approved the final version to be published. Dr. Eyre had full access to all of the data in the study and takes responsibility for the integrity of the data and the accuracy of the data analysis.

Study conception and design. Orozco, Wilson, Barton, Worthington, Eyre.

Acquisition of data. Orozco, Steer, Wordsworth, Hocking.

Analysis and interpretation of data. Orozco, Viatte, Bowes, Martin, Morgan, Barton, Eyre.

REFERENCES

1. Orozco G, Barton A. Update on the genetic risk factors for rheumatoid arthritis. Expert Rev Clin Immunol 2010;6:61–75.
2. Stahl EA, Raychaudhuri S, Remmers EF, Xie G, Eyre S, Thomson BP, et al. Genome-wide association study meta-analysis identifies seven new rheumatoid arthritis risk loci. Nat Genet 2010;42:508–14.
3. Barrett JC, Clayton DG, Concannon P, Akolkar B, Cooper JD, Erlich HA, et al. Genome-wide association study and meta-analysis find that over 40 loci affect risk of type 1 diabetes. Nat Genet 2009;41:703–7.
4. Anderson CA, Boucher G, Lees CW, Franke A, D'Amato M, Taylor KD, et al. Meta-analysis identifies 29 additional ulcerative colitis risk loci, increasing the number of confirmed associations to 47. Nat Genet 2011;43:246–52.
5. Franke A, McGovern DP, Barrett JC, Wang K, Radford-Smith GL, Ahmad T, et al. Genome-wide meta-analysis increases to 71 the number of confirmed Crohn's disease susceptibility loci. Nat Genet 2010;42:1118–25.
6. Wellcome Trust Case Control Consortium. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature 2007;447:661–78.
7. Howie BN, Donnelly P, Marchini J. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. PLoS Genet 2009;5:e1000529.
8. Li Y, Willer CJ, Ding J, Scheet P, Abecasis GR. MaCH: using sequence and genotype data to estimate haplotypes and unobserved genotypes. Genet Epidemiol 2010;34:816–34.
9. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet 2007;81:559–75.
10. Martin P, Barton A, Eyre S. ASSIMILATOR: a new tool to inform selection of associated genetic variants for functional studies. Bioinformatics 2011;27:144–6.
11. Zhernakova A, Stahl EA, Trynka G, Raychaudhuri S, Festen EA, Franke L, et al. Meta-analysis of genome-wide association studies in celiac disease and rheumatoid arthritis identifies fourteen non-HLA shared loci. PLoS Genet 2011;7:e1002004.
12. ENCODE Project Consortium. The ENCODE (ENCyclopedia Of DNA Elements) Project. Science 2004;306:636–40.
13. Imielinski M, Baldassano RN, Griffiths A, Russell RK, Annese V, Dubinsky M, et al. Common variants at five new loci associated with early-onset inflammatory bowel disease. Nat Genet 2009;41:1335–40.
14. Mells GF, Floyd JA, Morley KI, Cordell HJ, Franklin CS, Shin SY, et al. Genome-wide association study identifies 12 new susceptibility loci for primary biliary cirrhosis. Nat Genet 2011;43:329–32.
15. Sawcer S, Hellenthal G, Pirinen M, Spencer CC, Patsopoulos NA, Moutsianas L, et al. Genetic risk and a primary role for cell-mediated immune mechanisms in multiple sclerosis. Nature 2011;476:214–9.
16. Viatte S, Plant D, Bowes J, Lunt M, Eyre S, Barton A, et al. Genetic markers of rheumatoid arthritis susceptibility in anti-citrullinated peptide antibody negative patients. Ann Rheum Dis 2012;71:1984–90.

Publication 3: Enrichment of vitamin D response elements in RA-associated loci supports a role for vitamin D in the pathogenesis of RA

Genes and Immunity (2013) **14,** 325–329
© 2013 Macmillan Publishers Limited All rights reserved 1466-4879/13
www.nature.com/gene

ORIGINAL ARTICLE

Enrichment of vitamin D response elements in RA-associated loci supports a role for vitamin D in the pathogenesis of RA

A Yarwood[1,3], P Martin[1,3], J Bowes[1], M Lunt[1], J Worthington[1,2], A Barton[1,2] and S Eyre[1]

The aim of this study was to explore the role of vitamin D in rheumatoid arthritis (RA) pathogenesis by investigating the enrichment of vitamin D response elements (VDREs) in confirmed RA susceptibility loci and testing variants associated with vitamin D levels for association with RA. Bioinformatically, VDRE genomic positions were overlaid with non-HLA (human leukocyte antigen)-confirmed RA susceptibility regions. The number of VDREs at RA loci was compared to a randomly selected set of genomic loci to calculate an average relative risk (RR). Single-nucleotide polymorphisms (SNPs) in the *DHCR7/NADSYN1* (nicotinamide adenine dinucleotide synthase 1) and *CYP2R1* loci, previously associated with circulating vitamin D levels, were tested in UK RA cases ($n = 3870$) and controls ($n = 8430$). Significant enrichment of VDREs was seen at RA loci ($P = 9.23 \times 10^{-8}$) when regions were defined either by gene (RR 5.50) or position (RR 5.86). SNPs in the *DHCR7/NADSYN1* locus showed evidence of positive association with RA, rs4944076 ($P = 0.008$, odds ratio (OR) 1.14, 95% confidence interval (CI) 1.03–1.24). The significant enrichment of VDREs at RA-associated loci and the modest association of variants in loci-controlling levels of circulating vitamin D supports the hypothesis that vitamin D has a role in the development of RA.

Genes and Immunity (2013) **14,** 325–329; doi:10.1038/gene.2013.23; published online 2 May 2013

Keywords: vitamin D; rheumatoid arthritis; VDRE; Mendelian randomization

INTRODUCTION

Vitamin D is a steroid hormone that has an important role in many processes.[1] Vitamin D acts through its nuclear receptor, the vitamin D receptor, the expression of which on immune cells raises the question of a role for the hormone in the control of the immune system. Numerous lines of evidence support a regulatory role for vitamin D; indeed the active form, calcitriol or 1,25-dihydroxyvitamin D3 (VitD$_3$), has been found to control over 200 genes including those involved in cell differentiation, proliferation, apoptosis and angiogenesis.[2–4] The vitamin D receptor acts by binding to specific sequences in DNA known as vitamin D response elements (VDREs), which result in transcriptional regulation of vitamin D-responsive genes.

Rheumatoid arthritis (RA) is a chronic inflammatory autoimmune disease (AID) that causes inflammation of synovial joints. RA is thought to arise as a result of an autoimmune reaction due to a breach in self tolerance, inducing a type 1 T helper cell-driven immune response that causes infiltration of immune cells into the joint and secretion of pro-inflammatory cytokines.[5]

Vitamin D deficiency is common in RA and the active form of vitamin D can alter several aspects of the immune responses, which are pivotal to RA pathogenesis. The cytokine profile of immune cells can be affected; for example, production of interleukin (IL)-2 and interferon-γ, which are important immune mediators, by type 1 T helper lymphocytes is inhibited, resulting in a shift towards a regulatory type 2 T helper cell phenotype. Inhibition of interferon-γ production also reduces antigen presentation by antigen-presenting cells, and in turn reduces T-cell activation, acting as a negative feedback regulator.[6] Vitamin D inhibits the differentiation of monocytes to dendritic cells,[7] reduces the production of IL-12 by dendritic cells and stimulates phagacytosis of bacteria by macrophages.[8,9] Boonstra *et al.*[10] showed that vitamin D increased the production of IL-4, IL-5 and IL-10 by type 2 T helper cells; IL-4 and IL-10 normally act to inhibit type 1 T helper cell function. Vitamin D has also been shown to inhibit the activation of type 17 T helper cells by inhibiting the expression of IL-6.[11,12] In addition, vitamin D can affect the invasiveness of cultured RA fibroblast-like synoviocytes (FLS), in which higher concentrations of calcitriol were shown to significantly decrease the invasion of RA FLS by 53%.[13] The expression of matrix metalloproteinases (MMP), such as MMP-1 and MMP-2, can also be affected by vitamin D. MMPs are key molecules involved in the destruction of cartilage and bone in RA. IL-1β is known to increase the expression of MMP-1 by RA FLS; Laragione *et al.*[13] have shown that in both human and rats, treatment of activated FLS with calcitriol significantly inhibited IL-1β-induced expression of MMP-1 by 73–75%.

Given its immunoregulatory potential, a role for vitamin D has been postulated in AIDs. Circumstantial evidence supports this possibility; for example, circulating vitamin D levels have been reported to correlate with the risk of multiple sclerosis,[14] type 1 diabetes (T1D)[15] and Crohn's disease.[16] In RA, decreased serum levels of precursor vitamin D (25(OH)D) and active VitD$_3$ have been reported in cross-sectional studies.[17] However, findings may be confounded by the decreased physical activity and reduced sun exposure that RA patients with disability experience, which means that it is not clear whether the associations observed are cause or effect. With regards to disease onset, the

[1]Arthritis Research UK Epidemiology Unit, Centre for Musculoskeletal Research, Institute of Inflammation and Repair, The University of Manchester, Manchester Academic Health Science Centre, Manchester, UK and [2]NIHR Manchester Musculoskeletal Biomedical Research Unit, Manchester Academic Health Science Centre, Central Manchester Foundation Trust, Manchester, UK. Correspondence: Dr S Eyre, Arthritis Research UK Epidemiology Unit, Centre for Musculoskeletal Research, Institute of Inflammation and Repair, The University of Manchester, Manchester Academic Health Science Centre, Stopford Building, Oxford Road, Manchester M13 9PT, UK.
Email: Steve.eyre@manchetser.ac.uk

[3]These authors contributed equally to this manuscript.

Received 15 January 2013; revised 18 March 2013; accepted 21 March 2013; published online 2 May 2013

evidence surrounding vitamin D intake is conflicting. Merlino *et al.*[18] showed an inverse prospective association between high vitamin D intake and the risk of RA in women from the Iowa health study. However, a large prospective study of 180 000 women showed no association between vitamin D intake and subsequent development of RA.[19]

Low serum levels of vitamin D have also been associated with increased disease severity;[20–25] however, the regulatory genes associated with disease severity may be different to those associated with susceptibility, therefore, there may be a different set of genes regulating disease severity, which remain unexplored for VDRE enrichment.

Thus, there is conflicting epidemiological evidence for a role of vitamin D in influencing disease susceptibility to RA, and differentiating cause from effect is challenging. However, there is emerging genetic evidence that vitamin D levels may be linked to the onset of AIDs. First, a recent study by Ramagopalan *et al.*[26] determined VDREs throughout the genome using chromatin immunoprecipitation (ChIP), followed by massively parallel sequencing (ChIP-seq). The authors identified VDREs in lymphoblastoid cell lines from two individuals of European ancestry before and after VitD$_3$ stimulation. In the basal state, before stimulation, 623 genomic sites were identified, whereas after calcitriol stimulation, the number of VDRE sites increased to 2776. The authors demonstrated significant enrichment of VDREs in known AID susceptibility loci, including Crohn's disease, systemic lupus erythematosus, T1D and multiple sclerosis (*P*-value <0.001 in all diseases). Interestingly, the study also tested 16 RA susceptibility loci and found significant VDRE enrichment ($P<0.001$). However, only 9 of the 16 loci have been confirmed to be associated with RA; in addition, there are now 45 non-HLA (human leukocyte antigen) loci confirmed to be associated with RA susceptibility.[27]

The second piece of genetic evidence to support a role for vitamin D in the aetiology of AID comes from a recent study in a T1D cohort, this study confirmed the association of four vitamin D metabolism genes (*GC, DHCR7/NADSYN1, CYP2R1, CYP24A1*) with vitamin D levels in healthy controls; in addition, two of these genes showed association with T1D susceptibility (*DHCR7* $P=1.2\times10^{-3}$ and *CYP2R1* $P=3.0\times10^{-3}$).[28]

The association of the same variants in the same genes that control circulating levels of vitamin D with an AID provides an unbiased test of the hypothesis that vitamin D levels are important in the aetiology of AIDs (Mendelian randomization).[29] Therefore, the aims of the current study were, first, to investigate potential enrichment of VDREs in RA susceptibility loci that have been confirmed to date;[27] second, to test variants previously associated with circulating levels of vitamin D and T1D susceptibility for association with RA susceptibility.

RESULTS

Enrichment of VDREs at RA loci

Out of the 46 RA regions defined by gene, a total of 39 VDREs were identified in 17 RA regions, showing that 37% of RA regions contain VDREs. The relative risk (RR) for RA-associated genes harbouring VDRES compared with random genes was 5.5 (95% CI 2.55–26.33) (Table 1).

Analysing the data using a simple 2 × 2 table also showed a significant increase in VDREs in RA-associated loci compared with the rest of the protein-coding genes in the genome obtained from ensembl build 65, $P=1.20\times10^{-10}$ (odds ratio (OR) 5.72, 95% confidence interval (CI) 2.95–10.79) (Table 1).

However, this 2 × 2 table method does not take into account the number of VDREs in each gene region; therefore, a trend test was carried out using a 2 × 16 table (online Supplementary Table 2), which also showed a significant enrichment of VDREs ($P=9.23\times10^{-8}$).

When defining regions by gene, the RA-defined gene may not be the true causal gene. To overcome this issue, we also defined regions by position, extending 50 kb up- and downstream of the associated single-nucleotide polymorphism (SNP) site. Defining regions by SNP position identified 25 VDREs in 16 out of 46 regions (35.5% of RA-associated SNPs contain VDREs within 50 kb), again showing a significant enrichment of VDREs at RA loci (after 100 000 randomizations, RR 5.86, 95% CI 2.04–51) (Table 1). However, only one RA-associated SNP was shown to lie directly within a VDRE; rs947474 is located on chromosome 10 and maps within the *PRKCQ* locus.

Association of SNPs in vitamin D gene regions with RA

DHCR7/NADSYN1. After QC, 360 SNPs in 3870 UK RA cases and 8430 UK controls remained for analysis, capturing 59% of variation across the *DHCR7/NADSYN1* (nicotinamide adenine dinucleotide synthase 1) locus ($r^2>0.8$). Weak association ($P=0.04$) was seen at rs11600569, which is in complete linkage disequilibrium with the vitamin D SNP associated with T1D (rs12785878); the OR is the same in both diseases (OR 1.07) (Table 2).[28]

Association was seen between rs4944076, located in an intron of *NADSYN1* and RA ($P=0.008$, OR 1.14, 95% CI 1.03–1.24) after principal components analysis to correct for geographical variation (Figure 1, Table 2, full results Supplementary Table 3). This SNP was modestly correlated ($r^2=0.3$) with rs12785878.[28]

CYP2R1. After QC, 177 SNPs remained for analysis, capturing 47% of the variation across the *CYP2R1* region ($r^2>0.8$). Weak association was seen at four SNPs (Table 2): two of these SNPs, rs7116978 and rs6486205 ($P=0.05$ OR 1.06 95% CI 0.99:1.18), are highly correlated with rs10741657 and rs2060793 ($r^2=0.87$), previously associated with vitamin D levels and T1D[28] (Supplementary Figure 1).

DISCUSSION

Although RA loci have previously been shown to be enriched for VDREs, this was in a total of 16 non-confirmed loci, of which only nine have since been validated as RA susceptibility loci. Here, we perform the first comprehensive analysis of VDREs in all 45 confirmed non-HLA RA susceptibility loci. We have shown that genes or regions surrounding SNPs associated with susceptibility to RA are significantly enriched for VDREs. Indeed, the vitamin D receptor has a binding site at 17 out of 46 confirmed non-HLA

Table 1. VDRE enrichment results when regions are defined by gene or SNP position

	RR	*95% CI*	χ^2		
			P-value	*OR*	*95% CI*
Defining regions by gene	5.50	2.55–26.33	$1.21 \times 10 -^{10}$	5.72	2.95–10.79
Defining regions by position (50 kb)	5.86	2.04–51			

Abbreviations: CI, confidence interval; OR, odds ratio; RR, relative risk.

Table 2. Logistic regression results showing top SNP associations in the *DHCR7/NADSYN1 and CYP2R1* regions with RA

SNP	*SNP position*	*SNP type*	*MAF in controls*	*MAF in cases*	P-*value*	*OR*	*95% CI*
DHCR7/NADSYN1							
rs4944076	70889302	Intronic	0.09	0.10	0.008	1.135	1.034:1.246
rs4944997	70884016	Intronic	0.09	0.10	0.008	1.135	1.033:1.246
rs2919722	70631179	Intronic	0.40	0.42	0.014	1.072	1.014:1.133
rs1002171	70895219	Intronic	0.05	0.05	0.014	1.166	1.031:1.319
rs11600569	70851395	Intronic	0.22	0.23	0.040	1.071	1.003:1.143
CYP2R1							
rs117162870	14782929	Intronic	0.04	0.05	0.04	1.15	1.009:1.305
rs116856365	14726866	Intronic	0.04	0.05	0.04	1.42	1.004:1.300
rs7116978	14838347	Intronic	0.37	0.38	0.05	1.06	0.9996:1.118
rs6486205	14837832	Intronic	0.37	0.38	0.05	1.06	0.999:1.118

Abbreviations: CI, confidence interval; MAF, minor allele frequency; *NADSYN1,* nicotinamide adenine dinucleotide synthase 1; OR, odds ratio. rs11600569 shown in bold is perfectly correlated ($r^2 = 1$) with rs12785878 previously associated with vitamin D levels and T1D.[28]

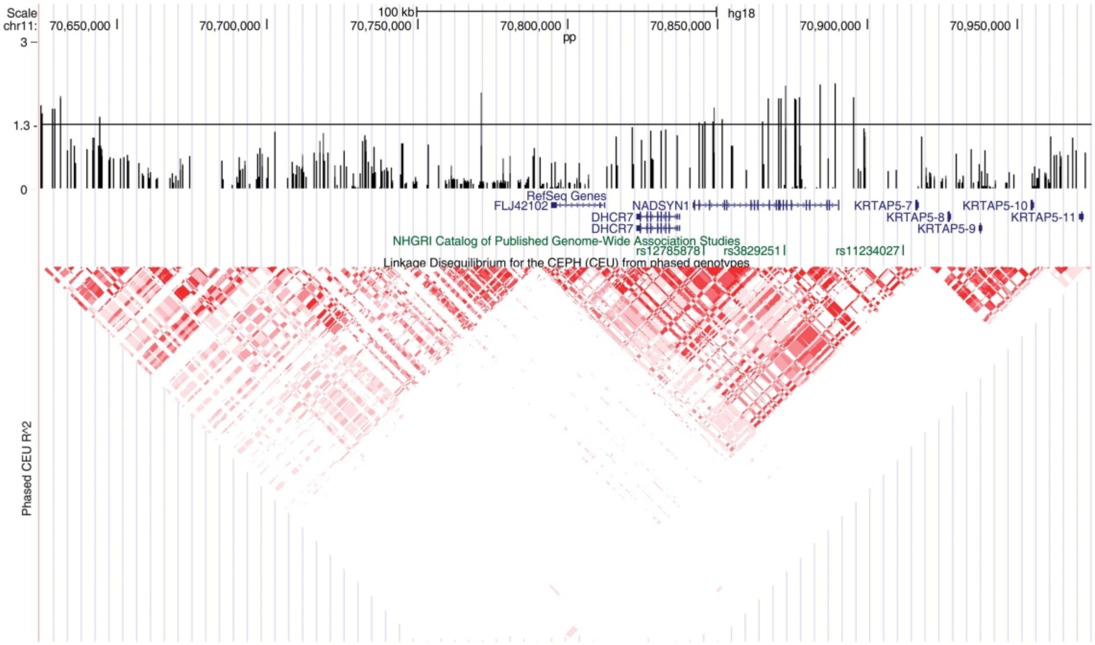


Figure 1. Fine mapping results showing the $-\log_{10}$ *P*-values of SNPs in the DHCR7/NADSYN1 region associated with RA. The *P*-value threshold is set at 1.3, which is equivalent to a *P*-value of 0.05. Reference sequence genes are shown in blue, HapMap LD is shown in red. SNPs in the region previously associated with circulating levels of vitamin D through GWAS are shown in green.

RA-associated genes, and binds within 50 kb of 35.5% of RA-associated SNPs, resulting in a RR of 5.86 compared with the rest of the genome. We have also shown that SNPs within the vitamin D-associated loci *DHCR7/NADSYN1* show modest association with RA susceptibility, suggesting that vitamin D may have a role in the development of RA.

There are some points that should be considered when interpreting these results. Primarily, the RA-associated genes as defined by this study may not represent the true causal gene. Indeed, when genes are assigned to loci, any functionality that is associated is largely speculative, and involves variants being assigned to the nearest gene, or to a gene in the region that is a plausible biological candidate. Therefore, we used GRAIL (gene relationships across implicated loci), which is the best method to assign genes to loci in a non-biased fashion. We also used a position-based approach, as well as a gene-based approach, with the assumption that a causal gene will lie reasonably close to the currently defined associated variant. However, associated variants may not exert their effects on nearby genes, but may be acting on genes some distance away in *cis* or *trans*.[30]

Although most regions have been fine-mapped using the immunochip (a custom SNP array designed for dense genotyping of 186 loci identified through genome-wide association studies),[27] in large sample collections, it is possible that the true causal variants are not the currently defined RA-associated variants or those in tight linkage disequilibrium with them and, indeed, independent associations may exist in already confirmed susceptibility loci. Therefore, more causal variants may exist within VDREs than indicated here, and functional studies to define causal SNPs will be necessary to understand the full picture.

Finally, it is possible that some genes and SNPs included in the randomized control groups will ultimately be associated with RA, once all contributing genetic factors have been identified. However, that would reduce the likelihood of a difference being observed between currently defined RA loci and random non-RA-associated loci, and would be more likely to lead to a type 2 error.

Reassuringly, results from the two methods to define RA regions produced very similar results. In addition, defining region by position showed an increased RR compared with defining regions by gene (RR 5.86 and RR 5.50, respectively) (Table 1), suggesting that as the regions are more specifically centred on the RA-associated variant, VDRE enrichment increases.

Several studies have found reduced vitamin D levels in individuals with RA and other AIDs,[14,16,17,19] but determining whether their association with disease is cause or effect is challenging. However, several SNPs have been reproducibly associated with vitamin D levels:[28,31] if the same SNPs are also associated with disease, it provides strong evidence that the pathway is causal (Mendelian randomization).[29] In the current study, the genetic results are of borderline significance; a SNP perfectly correlated ($r^2=1$) with the T1D-associated variant showed modest association (rs11600569 $P=0.04$) after correction for geographical variation. Although this association would not remain significant after correction for multiple testing, it is interesting to note that the effect size seen for the association of this SNP with RA is the same as was previously identified in T1D (OR 1.07),[28] increasing the plausibility of this result. In addition, we have shown association between other SNPs in the *DHCR7/NADSYN1* locus and RA ($P=0.008$), and although these SNPs have not been previously associated with vitamin D levels, since the association between SNPs in this region and vitamin D levels was identified by genome-wide association studies, the causal variant in the region responsible for altering vitamin D levels has not yet been determined, and it is possible that a secondary association in the region could be identified.

All variants associated with vitamin D levels (Table 2) are intronic variants. The most associated variant rs4944076 lies in an intron of NADSYN1. Bioinformatics analysis using the programme ASSIMILATOR[32] that retrieves and queries the experimental data generated by ENCODE, which is available on the UCSC web browser, has shown that rs4944076 lies within a region of open chromatin, DNase 1 hypersensitivity and histone modification, suggesting possible regulatory potential. In addition, one SNP, rs4944997, has been found to be an expression-quantitative-trait locus, potentially regulating the expression of NADSYN1.[33] However, the associations at this locus are modest and will require replication and identification of the functional variant before speculation of the functional effect of the variant can be determined.

The main source of vitamin D comes from endogenous production in the skin after exposure to UVB light from the sun, which results in the conversion of 7-dehydrocholesterol (present in the skin) to pre-vitamin D_3. This is then metabolized in the liver by CYP2R1 and CYP27A1 enzymes to form 25(OH)D, the major circulating form of vitamin D that is converted to its active form 1,25-(OH)2D3 (calcitriol) in the kidneys by CYP27B1. *DHCR7* encodes an enzyme that catalyzes the conversion of 7-dehydroxycholesterol to cholesterol, which removes pro-cholesterol from the vitamin D pathway, reducing the availability of 7-dehydrocholesterol for conversion to 25(OH)D. We could speculate that variants resulting in increased DHCR7 activity could, therefore, lead to increased removal of 7-dehydrocholesterol from the vitamin D pathway, causing vitamin D deficiency.

In conclusion, analysis of all 45 confirmed RA non-HLA susceptibility loci to date has shown a significant enrichment of VDREs at RA loci; in addition, we have shown the modest association of SNPs, previously associated with vitamin D levels, with RA. Validation of this finding is required in larger, independent studies. If confirmed, the genetic association along with the significant enrichment of VDREs in RA-associated loci would provide supportive evidence for the involvement of vitamin D in RA, and may pave the way for future trials of vitamin D therapy to prevent RA.

MATERIALS AND METHODS

Analysis of enrichment of VDREs at RA loci

The enrichment of VDREs at confirmed RA loci[27] was investigated bioinformatically to compare the genomic location of defined RA loci with the genomic positions of VDREs identified by Ramagopalan *et al.*[26] A total of 2776 VDREs identified by ChIP-seq in lymphoblastoid cell lines after VitD_3 stimulation were obtained from Ramagopalan *et al.*, and were assigned to the nearest gene within 100 kb using the Ensembl genome browser (www.ensembl.org) build 65. The 100 kb threshold was selected arbitrarily; it is possible that VDREs act on more distant genes, however it has been shown that calcitriol-responsive genes have a VDRE at a median distance of 66.6 Kb from the transcription start site.

RA loci were defined using the ensembl genome sequence in two ways:

(1) First, regions were defined based on the most plausible candidate gene, a query gene list was created using the 45 non-HLA loci defined in a recent study by Eyre *et al.*[27] (Supplementary Table 1). Each associated SNP from this study was assigned to the most functionally plausible candidate gene using GRAIL.[34] GRAIL uses the literature to identify relationships between genes, and selects the best candidate gene in a region in relation to a particular phenotype. One locus contained multiple candidate genes (*IL2–IL21*), and therefore was included twice, creating a final list of 46 genes in the query list.

(2) Second, regions were defined using the base pair position of each associated SNP from Eyre *et al.*[27] and defining the RA loci by extending 50 kb up- and downstream of each of the associated SNP (one locus IKZF3 had two candidate SNPs, see Supplementary Table 1).

The genomic locations of these defined RA loci obtained from Ensembl were then compared with the genomic positions of the VDREs from Ramagopalan *et al.*[26] using perl scripts to determine the number of VDREs present in RA loci.

To identify an enrichment of VDREs, a comparison set of the same number of random loci was created. Perl scripts were used to select random genes from all protein-coding genes in the genome (except those already associated with RA) or base pair positions from the ensembl genome sequence, and regions were defined in the same way as described for the RA loci. In total, 100 000 randomly selected comparison gene sets were generated and VDRE enrichment was determined by assessing the RR. This was calculated by dividing the number of VDREs in the query gene list by the average number of VDREs in 100 000 randomizations. Some random gene sets will undoubtedly contain zero VDREs, but the risk ratio cannot be calculated if there are no events in the control group. To overcome this problem, 0.5 was added to the total number of VDREs in all gene sets, including the query list.

STATA version 11 (http://www.stata.com) was then used to calculate *P*-values. A χ^2 test was performed to test the null hypothesis that the probability of a VDRE existing at a given locus did not depend on whether that locus was close to a region associated with RA risk, by creating a 2 × 2 table cross-classifying genes by association with RA and presence of a VDRE. As this method does not take into account the number of VDREs in each region, a trend test was performed using a 2 × N table.

Analysis of SNPs associated with circulating levels of vitamin D

The regions surrounding two genes previously associated with circulating levels of vitamin D and T1D were fine-mapped as part of a larger study[27] (*DHCR7/NADSYN1* and *CYP2R1*).[28,31] All known SNPs in the region from HapMap and the 1000 genomes project low-coverage whole-genome-sequencing pilot were included for fine mapping. Genotyping was carried out using a custom Illumina infinium genotyping chip, in 4752 UK RA cases and 9006 UK RA controls (described previously[27]).

As the *DHCR7/NADSYN1* locus lies in a region identified by the Wellcome Trust Case Control Consortium (WTCCC)[35] to be a region in which allele frequencies show geographic differentiation in the UK, principal components analysis[36] was carried out (described previously[27]) to correct for this. The study was approved by the North West Ethics Committee (MREC 99/8/84).

CONFLICT OF INTEREST

The authors declare no conflict of interest.

ACKNOWLEDGEMENTS

We thank Arthritis Research UK for their support (grant ref 17552). This work was supported by the innovative medicines initiative joint undertaking (IMI JU) funded project BeTheCure, (contract number 115142-2).

AUTHOR CONTRIBUTIONS

Conception and design of study: AY, SE, AB; acquisition and analysis of data: AY, PM, ML, JB; interpretation of data: AY, PM; manuscript preparation: AY, PM, SE, AB, JW.

REFERENCES

1 Holick MF. Vitamin D deficiency. *N Engl J Med* 2007; **357**: 266–281.

2 Holick MF. Resurrection of vitamin D deficiency and rickets. *J Clin Invest* 2006; **116**: 2062–2072.

3 DeLuca HF. Overview of general physiologic features and functions of vitamin D. *Am J Clin Nutr* 2004; **80**(Suppl 6): 1689S–1696S.

4 Holick MF. High prevalence of vitamin D inadequacy and implications for health. *Mayo Clin Proc* 2006; **81**: 353–373.

5 McInnes IB, O'Dell JR. State-of-the-art: rheumatoid arthritis. *Ann Rheum Dis* 2010; **69**: 1898–1906.

6 Cippitelli M, Santoni A. Vitamin D3: a transcriptional modulator of the interferon-gamma gene. *Eur J Immunol* 1998; **28**: 3017–3030.

7 DeLuca HF, Cantorna MT. Vitamin D: its role and uses in immunology. *FASEB J* 2001; **15**: 2579–2585.

8 Griffin MD, Lutz WH, Phan VA, Bachman LA, McKean DJ, Kumar R. Potent inhibition of dendritic cell differentiation and maturation by vitamin D analogs. *Biochem Biophys Res Commun* 2000; **270**: 701–708.

9 Griffin MD, Lutz W, Phan VA, Bachman LA, McKean DJ, Kumar R. Dendritic cell modulation by 1alpha,25 dihydroxyvitamin D3 and its analogs: a vitamin D receptor-dependent pathway that promotes a persistent state of immaturity *in vitro* and *in vivo*. *Proc Natl Acad Sci USA* 2001; **98**: 6800–6805.

10 Boonstra A, Barrat FJ, Crain C, Heath VL, Savelkoul HFJ, O'Garra A. 1{alpha}, 25-Dihydroxyvitamin D3 has a direct effect on naive CD4 + T cells to enhance the development of Th2 cells. *J Immunol* 2001; **167**: 4974–4980.

11 Stockinger B. Th17 cells: an orphan with influence. *Immunol Cell Biol* 2007; **85**: 83–84.

12 Xue ML, Zhu H, Thakur A, Willcox M. 1 alpha,25-Dihydroxyvitamin D3 inhibits pro-inflammatory cytokine and chemokine expression in human corneal epithelial cells colonized with *Pseudomonas aeruginosa*. *Immunol Cell Biol* 2002; **80**: 340–345.

13 Laragione T, Shah A, Gulko PS. The vitamin D receptor regulates rheumatoid arthritis synovial fibroblast invasion and morphology. *Mol Med* 2012; **18**: 194–200.

14 Ascherio A, Munger KL, Simon KC. Vitamin D and multiple sclerosis. *Lancet Neurol* 2010; **9**: 599–612.

15 Mathieu C, Gysemans C, Giulietti A, Bouillon R. Vitamin D and diabetes. *Diabetologia* 2005; **48**: 1247–1257.

16 Pappa HM, Grand RJ, Gordon CM. Report on the vitamin D status of adult and pediatric patients with inflammatory bowel disease and its significance for bone health and disease. *Inflamm Bowel Dis* 2006; **12**: 1162–1174.

17 Als OS, Riis B, Christiansen C. Serum concentration of vitamin D metabolites in rheumatoid arthritis. *Clin Rheumatol* 1987; **6**: 238–243.

18 Merlino LA, Curtis J, Mikuls TR, Cerhan JR, Criswell LA, Saag KG. Vitamin D intake is inversely associated with rheumatoid arthritis: results from the Iowa Women's Health Study. *Arthritis Rheum* 2004; **50**: 72–77.

19 Costenbader KH, Feskanich D, Holmes M, Karlson EW, ito-Garcia E. Vitamin D intake and risks of systemic lupus erythematosus and rheumatoid arthritis in women. *Ann Rheum Dis* 2008; **67**: 530–535.

20 Haga HJ, Schmedes A, Naderi Y, Moreno AM, Peen E. Severe deficiency of 25-hydroxyvitamin D(3) (25-OH-D (3)) is associated with high disease activity of rheumatoid arthritis. *Clin Rheumatol* (e-pub ahead of print 15 January 2013; doi:10.1007/s10067-012-2154-6).

21 Kostoglou-Athanassiou I, Athanassiou P, Lyraki A, Raftakis I, Antoniadis C. Vitamin D and rheumatoid arthritis. *Ther Adv Endocrinol Metab* 2012; **3**: 181–187.

22 Baykal T, Senel K, Alp F, Erdal A, Ugur M. Is there an association between serum 25-hydroxyvitamin D concentrations and disease activity in rheumatoid arthritis? *Bratisl Lek Listy* 2012; **113**: 610–611.

23 Higgins MJ, Mackie SL, Thalayasingam N, Bingham SJ, Hamilton J, Kelly CA. The effect of vitamin D levels on the assessment of disease activity in rheumatoid arthritis. *Clin Rheumatol* (e-pub ahead of print 23 January 2013; doi:10.1007/s10067-013-2174-x).

24 Song GG, Bae SC, Lee YH. Association between vitamin D intake and the risk of rheumatoid arthritis: a meta-analysis. *Clin Rheumatol* 2012; **31**: 1733–1739.

25 Rossini M, Maddali BS, La MG, Minisola G, Malavolta N, Bernini L *et al.* Vitamin D deficiency in rheumatoid arthritis: prevalence, determinants and associations with disease activity and disability. *Arthritis Res Ther* 2010; **12**: R216.

26 Ramagopalan SV, Heger A, Berlanga AJ, Maugeri NJ, Lincoln MR, Burrell A *et al.* A ChIP-seq defined genome-wide map of vitamin D receptor binding: associations with disease and evolution. *Genome Res* 2010; **20**: 1352–1360.

27 Eyre S, Bowes J, Diogo D, Lee A, Barton A, Martin P *et al.* High-density genetic mapping identifies new susceptibility loci for rheumatoid arthritis. *Nat Genet* 2012; **44**: 1336–1340.

28 Cooper JD, Smyth DJ, Walker NM, Stevens H, Burren OS, Wallace C *et al.* Inherited variation in vitamin D genes is associated with predisposition to autoimmune disease type 1 diabetes. *Diabetes* 2011; **60**: 1624–1631.

29 Hingorani A, Humphries S. Nature's randomised trials. *Lancet* 2003; **366**: 1906–1908.

30 Pomerantz MM, Ahmadiyeh N, Jia L, Herman P, Verzi MP, Doddapaneni H *et al.* The 8q24 cancer risk variant rs6983267 shows long-range interaction with MYC in colorectal cancer. *Nat Genet* 2009; **41**: 882–884.

31 Wang TJ, Zhang F, Richards JB, Kestenbaum B, van Meurs JB, Berry D *et al.* Common genetic determinants of vitamin D insufficiency: a genome-wide association study. *Lancet* 2010; **376**: 180–188.

32 Martin P, Barton A, Eyre S. ASSIMILATOR: a new tool to inform selection of associated genetic variants for functional studies. *Bioinformatics* 2011; **27**: 144–146.

33 Veyrieras JB, Kudaravalli S, Kim SY, Dermitzakis ET, Gilad Y, Stephens M *et al.* High-resolution mapping of expression-QTLs yields insight into human gene regulation. *PLoS Genet* 2008; **4**: e1000214.

34 Raychaudhuri S, Plenge RM, Rossin EJ, Ng AC, Purcell SM, Sklar P *et al.* Identifying relationships among genomic disease regions: predicting genes at pathogenic SNP associations and rare deletions. *PLoS Genet* 2009; **5**: e1000534.

35 The Wellcome Trust Case Control Consortium. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* 2007; **447**: 661–678.

36 Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D. Principal components analysis corrects for stratification in genome-wide association studies. *Nat Genet* 2006; **38**: 904–909.

Supplementary Information accompanies this paper on Genes and Immunity website (http://www.nature.com/gene)

Publication 4: Capture Hi-C reveals novel candidate genes and complex long-range interactions with related autoimmune risk loci

nature COMMUNICATIONS

ARTICLE

Received 15 Jun 2015 | Accepted 28 Oct 2015 | Published 30 Nov 2015

DOI: 10.1038/ncomms10069

OPEN

Capture Hi-C reveals novel candidate genes and complex long-range interactions with related autoimmune risk loci

Paul Martin[1,*], Amanda McGovern[1,*], Gisela Orozco[1,*], Kate Duffus[1], Annie Yarwood[1], Stefan Schoenfelder[2], Nicholas J. Cooper[3], Anne Barton[1,4], Chris Wallace[3,5], Peter Fraser[2], Jane Worthington[1,4] & Steve Eyre[1]

Genome-wide association studies have been tremendously successful in identifying genetic variants associated with complex diseases. The majority of association signals are intergenic and evidence is accumulating that a high proportion of signals lie in enhancer regions. We use Capture Hi-C to investigate, for the first time, the interactions between associated variants for four autoimmune diseases and their functional targets in B- and T-cell lines. Here we report numerous looping interactions and provide evidence that only a minority of interactions are common to both B- and T-cell lines, suggesting interactions may be highly cell-type specific; some disease-associated SNPs do not interact with the nearest gene but with more compelling candidate genes (for example, *FOXO1*, *AZI2*) often situated several megabases away; and finally, regions associated with different autoimmune diseases interact with each other and the same promoter suggesting common autoimmune gene targets (for example, *PTPRC*, *DEXI* and *ZFP36L1*).

[1] Arthritis Research UK Centre for Genetics and Genomics, Centre for Musculoskeletal Research, Institute of Inflammation and Repair, Faculty of Medical and Human Sciences, Manchester Academic Health Science Centre, The University of Manchester, Stopford Building, Oxford Road, Manchester M13 9PT, UK. [2] Nuclear Dynamics Programme, The Babraham Institute, Cambridge CB22 3AT, UK. [3] JDRF/Wellcome Trust Diabetes and Inflammation Laboratory, Department of Medical Genetics, NIHR Cambridge Biomedical Research Centre, Cambridge Institute for Medical Research, University of Cambridge, Wellcome Trust/MRC Building, Cambridge Biomedical Campus, Cambridge CB2 0XY, UK. [4] NIHR Manchester Musculoskeletal Biomedical Research Unit, Central Manchester Foundation Trust, Manchester Academic Health Science Centre, Oxford Road, Manchester M13 9WL, UK. [5] MRC Biostatistics Unit, Cambridge Institute of Public Health, Forvie Site, Robinson Way, Cambridge Biomedical Campus, Cambridge CB2 0SR, UK. * These authors contributed equally to this work. Correspondence and requests for materials should be addressed to S.E. (email: steve.eyre@manchester.ac.uk).

The idenfication of the precise gene targets of variants associated with complex traits detected through genome-wide association studies (GWAS) has proved challenging[1] but is essential if the full potential of genetic studies is to be realised. Accumulating evidence suggests the majority of these variants lie outside traditional protein-coding genes and are enriched in enhancer regions, which are both cell-type and stimulus specific[2–4]. The task now is to identify which genes are implicated and understand which cell types are involved, to ascertain the biological pathways that are perturbed in individuals who are genetically susceptible to disease. It is well-established that enhancers regulate gene transcription by physical interactions[5]. These can operate over large genetic distances, so the tradition of annotating GWAS hits with the closest, or most biologically plausible gene candidate, may prove misleading and result in expensive, time consuming efforts to define the function of non-causal genes.

The utility of chromosome conformation capture technology (Capture Hi-C) to detect the patterns of interactions between chromosomal regions has been demonstrated[6–9]. Here, for the first time, we used this approach to characterize the interactions of confirmed susceptibility loci for four autoimmune diseases: rheumatoid arthritis (RA), type 1 diabetes (T1D), psoriatic arthritis (PsA) and juvenile idiopathic arthritis (JIA) with the aim of linking disease-associated SNPs with disease-causing genes. Uniquely, we have tested the interactions in two complementary experiments: first, Region Capture targets regions associated with disease[10–14]; second, Promoter Capture provides independent validation through capturing all known promoters within 500 kb of lead disease-associated single nucleotide polymorphisms (SNPs). Our study expands on recent applications of the Capture Hi-C method firstly, by increasing the depth of sequencing and therefore the resolution, (average 10,000 interactions per restriction fragment), second, by comprehensively targeting the full known genetic component of four related autoimmune diseases and finally by performing complimentary experiments, such that we target the disease-associated regions and, in separate experiments, all gene promoters within 500 kb, so providing direct, independent, reciprocal validation for each interaction. All experiments were performed in human B (GM12878) and T (Jurkat) cell lines, selected because they are most relevant to these diseases[3]. Hi-C libraries were generated for both cell lines[15], then hybridized to custom biotinylated RNA baits and sequenced on an Illumina HiSeq 2500. We tested for significant interactions using a negative binomial distribution as described previously[6], performing all experiments in duplicate.

Our findings provide compelling evidence that disease-associated SNPs, currently nominally assigned to the closest plausible gene candidate, may well-regulate genes some distance away. We also show that in a subset of risk loci, SNPs associated to different autoimmune diseases physically interact with and may well-regulate the same genes but with differing enhancer mechanisms. A number of the interactions also show evidence of cell-type specificity, occurring in either the B- or T-cell lines only.

Results

Summary of identified interactions. Our unique study design determined a complex array of interactions between disease-associated regions and promoters (Fig. 1). After quality control, in the Region Capture experiment, 60.9 million and 54.9 million unique di-tags (comprising one restriction fragment from a capture target region and its ligated interacting partner) were on-target for GM12878 and Jurkat cell lines, respectively (average 21,170 reads per HindIII restriction fragment; 62% capture efficiency). Similarly, in the Promoter Capture experiment, 121.1 million (GM12878) and 115 million (Jurkat) unique di-tags were on-target (average 21,448 reads per HindIII restriction fragment; 70% capture efficiency) (Fig. 2).

At any given false discovery rate (FDR) threshold, interactions are called with an unknown rate of false negatives. With the assumption that interactions called in both the Region and Promoter Capture experiments are more likely to be true positives compared with those only seen in one experiment, we evaluated several potential FDR thresholds (Fig. 3). We saw a consistent

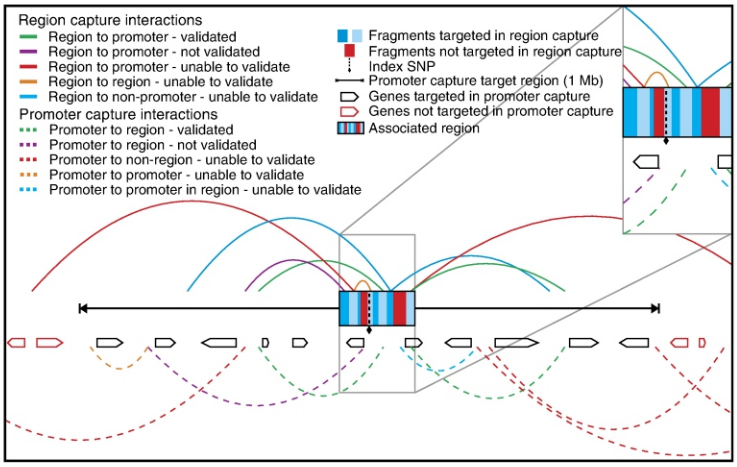

Region capture interactions
Region to promoter - validated
Region to promoter - not validated
Region to promoter - unable to validate
Region to region - unable to validate
Region to non-promoter - unable to validate
Promoter capture interactions
Promoter to region - validated
Promoter to region - not validated
Promoter to non-region - unable to validate
Promoter to promoter - unable to validate
Promoter to promoter in region - unable to validate
Fragments targeted in region capture
Fragments not targeted in region capture
Index SNP
Promoter capture target region (1 Mb)
Genes targeted in promoter capture
Genes not targeted in promoter capture
Associated region

Figure 1 | A schematic of a hypothetical associated region including possible chromatin interactions. Chromatin interactions are shown by arcs, those above the promoter capture target region are observed in the ‘Region Capture’ experiment; those below are observed in the ‘Promoter Capture’ experiment. All potential chromatin interactions are shown and are coloured by their potential to appear and be validated in both capture experiments. Those in green are observed in both the ‘Region Capture’ and the ‘Promoter Capture’ and comprise the ‘confirmed’ interaction set. Interactions shown in purple are only present in one capture experiment and were therefore not validated. Other interactions (red, orange and blue) would only be observed in either the ‘Region Capture’ or ‘Promoter Capture’ and could therefore not be validated as described. The inset shows a magnified view of the associated region (as defined by LD) detailing which restriction fragments were targeted in the ‘Region Capture’ and which were excluded as they appeared in the ‘Promoter Capture’.

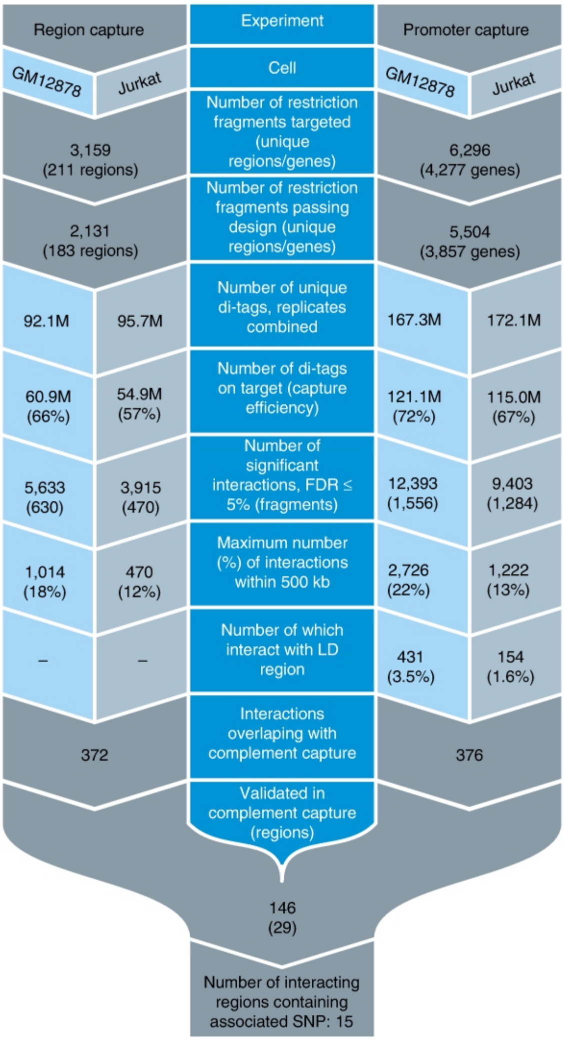

Region capture
Experiment
Promoter capture
GM12878
Jurkat
Cell
GM12878
Jurkat
3,159 (211 regions)
Number of restriction fragments targeted (unique regions/genes)
6,296 (4,277 genes)
2,131 (183 regions)
Number of restriction fragments passing design (unique regions/genes)
5,504 (3,857 genes)
92.1M
95.7M
Number of unique di-tags, replicates combined
167.3M
172.1M
60.9M (66%)
54.9M (57%)
Number of di-tags on target (capture efficiency)
121.1M (72%)
115.0M (67%)
5,633 (630)
3,915 (470)
Number of significant interactions, FDR ≤ 5% (fragments)
12,393 (1,556)
9,403 (1,284)
1,014 (18%)
470 (12%)
Maximum number (%) of interactions within 500 kb
2,726 (22%)
1,222 (13%)
–
–
Number of which interact with LD region
431 (3.5%)
154 (1.6%)
372
Interactions overlaping with complement capture
376
Validated in complement capture (regions)
146 (29)
Number of interacting regions containing associated SNP: 15

Figure 2 | Flowchart summarizing capture Hi-C experiments by cell line. The 'Region Capture' experiment is shown on the left and the 'Promoter Capture' experiment on the right. Flowchart sections are coloured by cell type: light blue—GM12878 cells; light grey—Jurkat cells and grey—both cell types. Each section label is shown in dark blue.

enrichment in interactions called in both experiments at decreasing Promoter Capture experiment FDR thresholds, providing confidence that they represent true interactions. At 5% FDR, we called 8,594 interactions in the Region Capture experiment representing 764 targeted HindIII restriction fragments. Of these interactions 372/8,594 (4.3%) from 116 targeted HindIII restriction fragments demonstrated evidence of interacting with a promoter within 500 kb, and so could be validated by the complementary capture method. Of these, 146/342 interactions were identified in the Promoter Capture experiment (Fig. 2), implicating 29 regions, of which 15 contain disease-associated SNPs (Supplementary Table 1). The majority of significant interactions were cell-type specific, with only 20% found in both cell lines.

We compared our data with publicly available chromatin interaction data in similar cell lines and could detect the well-established interactions with the *cis*-acting regulatory region of the *HBA* locus[16] (Supplementary Fig. 1a) and interactions in the 5C ENCODE (https://www.encodeproject.org/)[17] experiments at two regions: *IFNAR1* and *IL5* (Supplementary Fig. 1b,c).

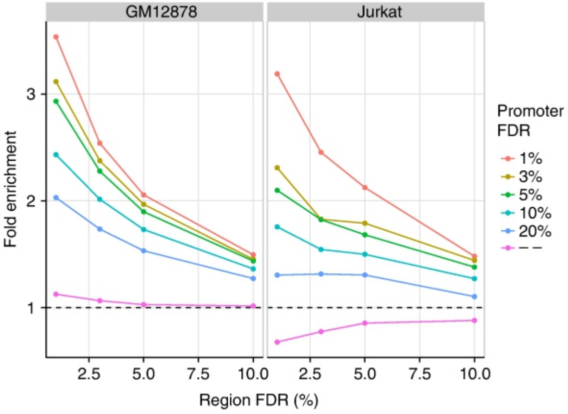


Figure 3 | Fold enrichment. Fold enrichment of retained interactions called in the promoter capture experiments with decreasing FDR thresholds, given they had been called in the region capture experiments at the FDR threshold shown. '—' shows the enrichment found by focusing only on interactions called in the region capture experiments for which the other end lay in a HindIII restriction fragment targeted in the promoter capture design.

Interactions with novel candidate genes. Confirmed interactions provided examples of disease-associated SNPs that do not interact with the nearest gene, but rather with promoters some distance away, implicating entirely different target genes. For example, strong evidence was found to suggest that regions with SNPs associated with RA, situated proximal to the *EOMES* gene, make strong physical contact with the promoter of *AZI2*, a gene involved in NFκB activation, some 640 kb away (Fig. 4a) in both GM12878 and Jurkat cell lines. In addition, variants associated with RA and JIA in the 3′ intronic region of *COG6,* a gene encoding a component of Golgi apparatus, show interactions with the promoter of the *FOXO1* gene, mapping over 1 Mb away, in both cell types (Fig. 4b). Recent findings suggest that the *FOXO1* gene is important in the survival of fibroblast-like synoviocytes (FLS) in RA[18] and is hypermethylated in RA FLS compared with osteoarthritis FLS[19], providing strong supporting functional evidence as to gene candidature.

Common interaction targets mediated by multiple genetic loci. Perhaps the most striking finding comes from genetic regions that harbour susceptibility loci for different autoimmune diseases, where the lead disease-associated SNP for one disease maps some distance from the lead disease-associated SNP for other autoimmune diseases; previously, using the 'nearest candidate gene' annotation method, different genes would have been assigned to the diseases but our work shows that they may all act on the same gene promoter. We provide three examples below to illustrate the findings. First, the 16p13 region contains SNPs associated with both T1D and multiple sclerosis that locate within intron 19 of the *CLEC16A* gene. A physical interaction between a 20-kb region of *CLEC16A* and the promoter of *DEXI* has previously been reported[20], although was not detected in the current study. Our data suggest that a separate, independent region, associated with both T1D and JIA, near the *RMI2* gene and 530 kb from the *DEXI* gene, also interacts with the *DEXI*

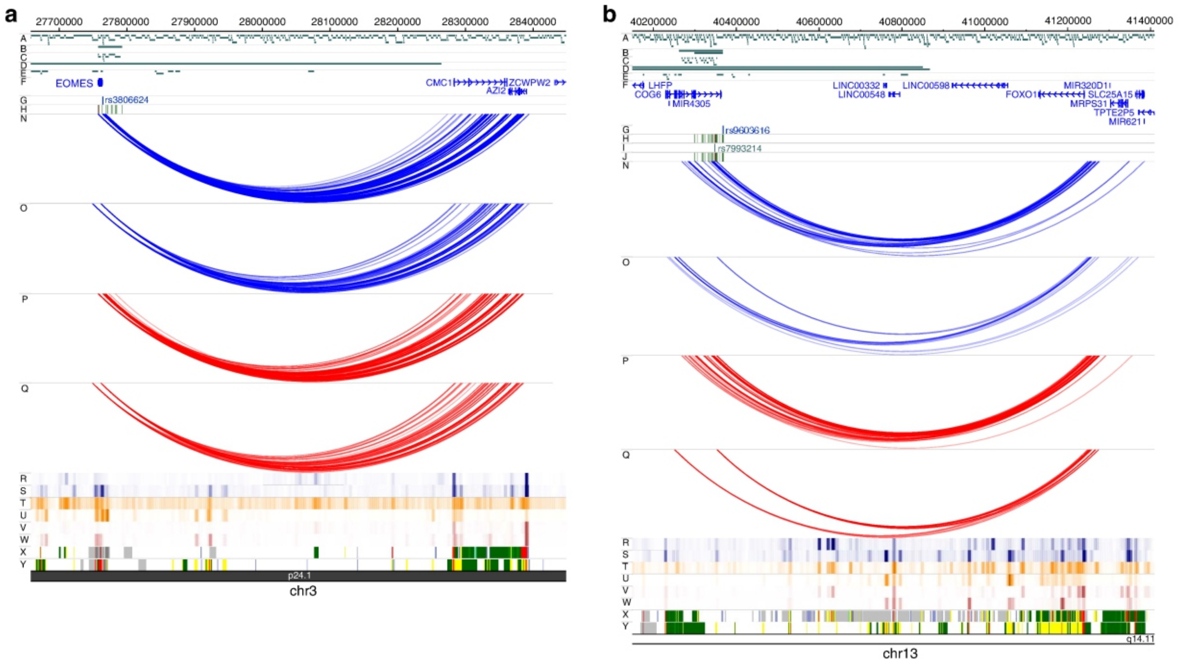


Figure 4 | Examples of chromatin interactions implicating novel gene candidates. (**a**) *EOMES* SNPs—both GM12878 and Jurkat cell lines show that SNPs situated proximal to the *EOMES* gene interact with the promoter of *AZI2I*, involved in NFκB activation, situated ~640 kb away. (**b**) *COG6* SNPs—interactions are shown that link SNPs within the *COG6* to the *FOXO1* promoter, over 1 Mb away, in both cell types. Genomic co-ordinates are shown along the top of each panel and tracks are labelled A–Y (empty tracks removed for clarity): (A) HindIII restriction fragments; (B–E) Regions targeted and restriction fragments included in the region (B,C) and promoter (D,E) capture experiments; (F) RefSeq Genes from the UCSC Genome Browser, downloaded 1 January 2012; (G,I,K) Index SNPs identified for RA (G), JIA (I) and PsA (K). Associations in red were identified in the RA Immunochip study. SNPs in blue were novel associations identified in the RA *trans*-ethnic GWAS meta-analysis, JIA and PsA SNPs were identified in the JIA and PsA Immunochip studies; (H,J,L) Density plots showing 1000 Genomes SNPs in LD ($r^2 \geq 0.8$) with the index SNPs (green-red) for RA (H), JIA (J) and PsA (L); (M) T1D Credible set SNPs identified in the T1D Immunochip study; (N–Q) Significant Interactions identified in the region and promoter capture experiments in GM12878 (N,O) and Jurkat (P,Q) cells; (R–Y) Data from the WashU Encode track hub showing DNaseI HS sites, H3K4me1 histone marks and H3K27ac histone marks for GM12878 (R,T,V) and CD3 Primary (S,U,W) cells and BROAD ChromHMM states for GM12878 (X) and CD4 Naive Primary cells (Y).

promoter (Fig. 5a). Furthermore, a region proximal to the *ZC3H7A* gene, associated with RA susceptibility, some 1.2 Mb from *DEXI*, interacts with both the T1D/JIA-associated region and the *DEXI* promoter.

The second example is provided by RA-associated variants mapping within a strong enhancer region intronic of *RAD51B*, where a significant interaction is observed with the promoter of the *ZFP36L1* gene. SNPs in the promoter region of *ZFP36L1* are independently associated with JIA but not RA; however, the interaction of the *ZFP36L1* promoter with the RA-associated SNPs suggests that the causal gene in both diseases may be *ZFP36L1* and not *RAD51B*. *ZFP36L1* is a zinc finger transcription factor involved in the transition of B cells to plasma cells and it is noteworthy that the interaction with the RA-associated region was only seen in the B-cell line (Fig. 5b).

Finally we show evidence that SNPs associated with PsA within the *DENND1B* gene make strong contact with a region associated with RA within the *PTPRC* gene, which is responsible for T- and B-cell receptor signalling and maps over 1 Mb away (Fig. 5c).

We, like others[8,9], have demonstrated a complex relationship between promoters and enhancers, where promoters interact with many enhancers and enhancers interact with many promoters, rarely in a one-to-one relationship (Fig. 1 and Supplementary Table 2). Enhancers containing risk variants for autoimmune diseases can, therefore, 'meet' at the same promoters. This challenges the assumption that disease-associated SNPs have to be in close linkage disequilibrium (LD) to have a disease related effect on the same gene. In addition, these findings may well-suggest an evolutionary phylogeny, where polymorphic variants regulating expression of the same gene result in either different autoimmune diseases or different molecular mechanisms resulting in risk of the same disease.

Interactions with previously implicated loci. Among the other 141 confirmed interactions, we observed examples of disease-associated SNPs within the 3′ untranslated region, or within introns of a gene, interacting with the promoter of the same gene (*STAT4*, *CDK6*, Supplementary Fig. 2a,b); disease-associated SNPs within lncRNA interacting with the promoter of genes (*RBPJ*, Supplementary Fig. 3) and several examples of restriction fragments, proximal to those containing disease-associated SNPs, interacting with promoters some distance away (*ARID5B*, *IL2RA*, *TLE3*, Supplementary Fig. 4a–c), supporting recent findings that disease-associated SNPs are enriched outside transcription factor-binding sites[3].

Long-range interactions. Perhaps unexpectedly, ~80% of significant interactions occurred at distances exceeding 500 kb (Supplementary Fig. 5) and interacted with 'non-promoters',

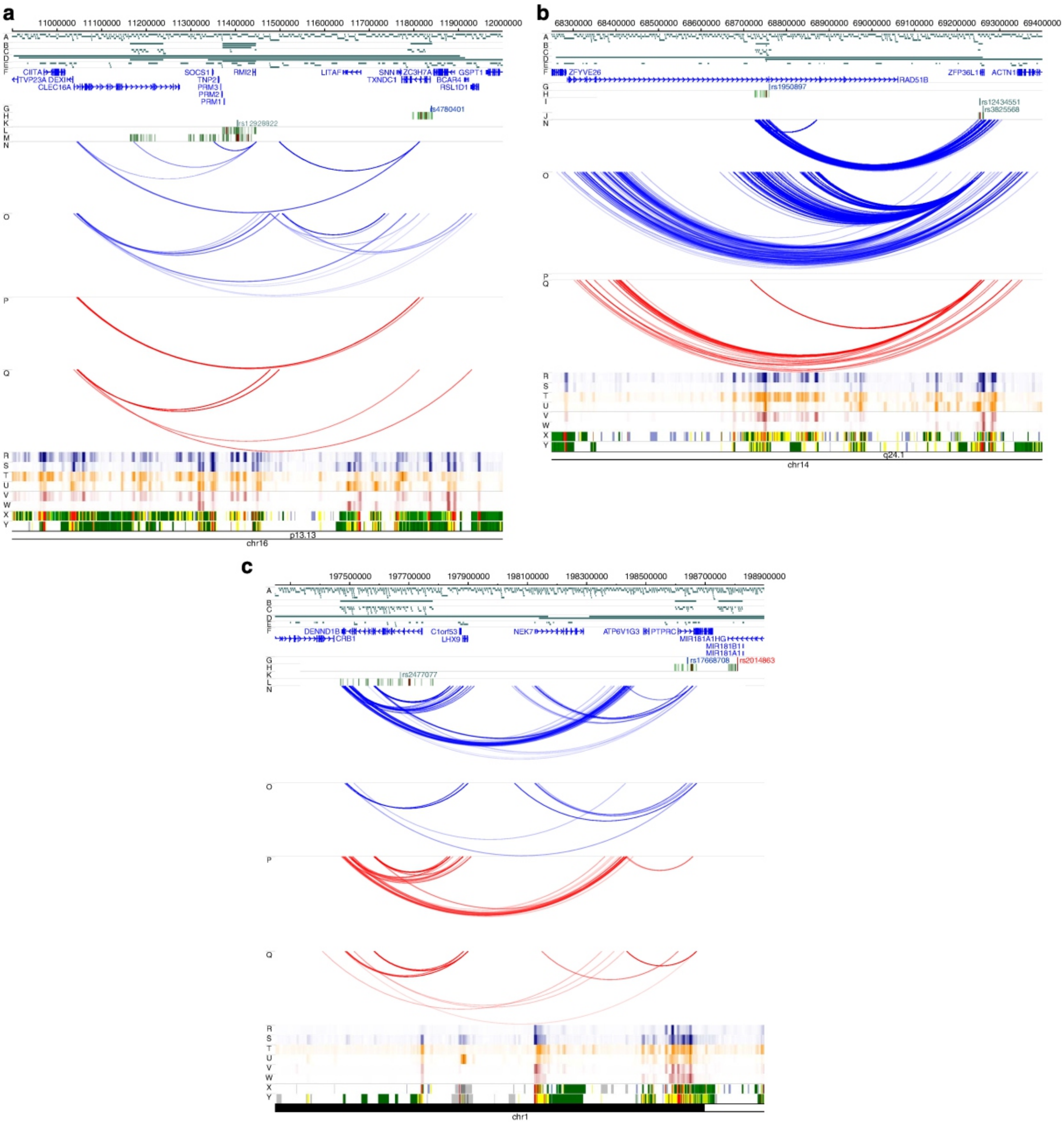


Figure 5 | Examples of chromatin interactions linking several disease associations to a common promoter. (**a**) *DEXI*—both GM12878 and Jurkat cell lines show that SNPs associated independently with RA, PsA and T1D interact with the *DEXI* promoter. In addition, evidence suggests that the RA and JIA SNP regions interact in GM12878 cells. (**b**) *RAD51B*—RA associations located within a strong enhancer are shown to interact with the promoter of *ZFP36L1*, a gene involved in B-cell transition, which also contains SNPs associated with JIA. (**c**), *PTPRC*—Variants associated with PsA, within the *DENND1B* are shown to interact with *PTPRC*, a region independently associated with RA. Genomic co-ordinates are shown along the top of each panel and tracks are labelled A—Y as in Fig. 4.

reducing the number of interactions available for co-validation in the Promoter and Region Capture experiments (targeted genes in the Promoter Capture not extending that far) and reinforcing the idea that GWAS regions may be involved with complex long-range gene regulation possibly involving multiple enhancer elements. To investigate whether these are likely to be true interactions, we compared results from the largest Hi-C data set on GM12878 cells reported, to date[21]. Of the 4,607 longer distance interactions (>500 kb) we called at FDR <5% in our data, 377 were found at 50 times observed over expected in the independent Hi-C data set (Supplementary Data 1). This provided both strong confirmation of our long-range capture Hi-C results we already co-validated with Promoter and Region Capture (for example, *FOXO1*, *ZFP36L1*, Supplementary Fig. 6) and supports many potentially novel interactions (for example, *MMEL1*, Supplementary Fig. 7), but detailed examination to confirm these long-range interactions is now required.

Discussion

Our targeted Capture Hi-C analyses have identified, for the first time, many long-range interactions between autoimmune risk loci and their putative target genes. Using this methodology we have intriguing data illustrating that regions associated with more than one disease, often some distance apart, interact with the same gene and that associated regions can 'skip' genes to interact with more distant novel candidates. Our results provide new insights into complex disease genetics and changes the way we view the causal genes in disease, with obvious implications for pathway analysis and identification of therapeutic targets. Since we uncovered evidence of cell-specific interactions, the current study is likely to be only the beginning of similar explorations. Further work to characterize functionally the observed interactions, including eQTL studies using a range of cell types and stimulatory conditions, are required to determine how disease-associated SNPs influence the risk of disease, with the aim of better understanding disease aetiology.

Methods

SNP and region associations. All independent lead disease-associated SNPs for RA were selected from both the fine-mapped Immunochip study[10] and a *trans*-ethnic GWAS meta-analysis[11]. Lead disease-associated SNPs were also added from the Immunochip fine mapping studies for JIA[13] and PsA[12]. This resulted in a total of 242 distinct variants associated with one or more of the three diseases after exclusion of *HLA*-associated SNPs. Associated regions were defined by selecting all SNPs in LD with the lead disease-associated SNP ($r^2> = 0.8$; 1000 Genomes phase 1 EUR samples; May 2011). In addition to the SNP associations, credible SNP set regions were defined for both T1D- and RA-associated loci discovered by the Immunochip array at a 99% confidence level[14]. RA loci, as defined from the Immunochip analysis, were extended to include the credible SNP region where necessary and overlapping regions were merged using the BEDTools v2.21.0 (ref. 22) merge command resulting in 211 associated regions.

Target enrichment design. To remain hypothesis free and to validate significant findings, two target enrichments were designed. The first targeted the 'associated region' and was called the 'Region Capture' set. The second targeted all known gene promoters overlapping the region 500 kb up- and downstream of the lead disease-associated SNP dubbed as the 'Promoter Capture' set. Capture oligos (120 bp; 25–65% GC, <3 unknown (N) bases) were designed using a custom Perl script within 400 bp but as close as possible to each end of the targeted HindIII restriction fragments and submitted to the Agilent eArray software (Agilent) for manufacture.

Region Capture design. Capture oligonucleotides were designed to all HindIII restriction fragments in each previously defined associated region after excluding those already targeted in the Promoter Capture. Regions were extended by one restriction fragment where there was <500 bp between the restriction site and the region start/end. This resulted in 3,159 restriction fragments in total after merging overlapping regions. Of these, 1,028 failed design, 1,096 had both ends captured and 1,035 had one end captured, producing a target capture of 387.24 kb covering a genomic region of 7.46 Mb (3.5 kb/restriction fragment on average). In addition, a control region, which represents a well-characterized region of long-range interactions, was also included: *HBA* (174.57 kb genomic; 26 restriction fragments; 6.71 kb/restriction fragment).

Promoter Capture design. Promoter Capture target regions were defined as 500 kb up- and downstream of each disease-associated SNP. These regions were further extended to encompass the associated regions where appropriate. HindIII restriction fragments were identified within 500 bp of the transcription start site of all genes mapping to the defined regions (Ensembl release 75; GRCh37) and overlapping regions were merged using the BEDTools[22] merge command resulting in 6,296 restriction fragments. Of these, 792 failed design, 2,986 had both ends captured and 2,518 had one end captured, producing a target capture of 1.02 Mb. The 5,504 captured restriction fragments covered a genomic region of 38.76 Mb (7.04 kb/restriction fragment on average) and contained promoters for 3,857 genes. The *HBA* control region previously mentioned was also included.

Cell culture and crosslinking. The GM12878 B-lymphoblastoid cell line, produced from the blood of a female donor with northern and western European ancestry by EBV transformation, was obtained from Coriell Institute for Medical Research. Lymphoblastoid cell lines were cultured in Roswell Park Memorial Institute (RPMI) 1640 per 20 mM L-glutamine supplemented with 15% foetal bovine serum (FBS) in 25 cm^2 vented culture flasks at 37 °C per 5% CO_2. The T-lymphoblastoid Jurkat E6.1 cell line, originating from the peripheral blood of a 14-year-old boy in the study by Schneider *et al.*[23], was obtained from LGC Standards and cultured in RPMI 1640 per 20 mM L-glutamine supplemented with 10% FBS in 25 cm^2 vented culture flasks at 37 °C/5% CO_2. To generate Hi-C libraries, $5–6 \times 10^7$ GM12878 and Jurkat cells were grown to $\sim 90\%$ confluence then formaldehyde crosslinking was carried out as described in the study by Belton *et al.*[15]. Cells were washed in Dulbecco's Modified Eagle's medium (DMEM) without serum then crosslinked with 2% formaldehyde for 10 min at room temperature. The crosslinking reaction was quenched by adding cold 1 M glycine to a final concentration of 0.125 M for 5 min at room temperature, followed by 15 min on ice. Crosslinked cells were washed in ice-cold PBS, the supernatant discarded and the pellets flash-frozen in liquid nitrogen and stored at -80 °C.

Hi-C library generation. Cells were thawed on ice and re-suspended in 50 ml freshly prepared ice-cold lysis buffer (10 mM Tris-HCl pH 8, 10 mM NaCl, 0.2% Igepal CA-630, one protease inhibitor cocktail tablet). Routinely, two pellets from each cell line were re-suspended and combined in 7 ml complete lysis buffer to give ~ 5-6×10^7 cells. Cells were lysed on ice for a total of 30 min, with 2×10 strokes of a Dounce homogeniser with a 5-min break between Douncing. Following lysis, the nuclei were pelleted and washed with $1.25 \times$ NEB Buffer 2 then re-suspended in $1.25 \times$ NEB Buffer 2 to make aliquots of $5–6 \times 10^6$ cells for digestion. Following lysis, Hi-C libraries were digested using HindIII then prepared as described in the study by van Berkum *et al.*[24] with modifications described in the study by Dryden *et al.*[6]. Pre-Capture amplification was performed with eight cycles of PCR on multiple parallel reactions from Hi-C libraries immobilized on Streptavidin beads, which were pooled post PCR and SPRI bead purified. The final library was re-suspended in 30 μl TLE and the quality and quantity assessed by Bioanalyzer and qPCR.

Solution hybridization capture of Hi-C library. Hi-C samples corresponding to 750 ng were concentrated in a Speedvac then re-suspended in 3.4 μl water. Hybridization of SureSelect custom Promoter and Region Capture libraries to Hi-C libraries was carried out using Agilent SureSelectXT reagents and protocols. Post-capture amplification was carried out using six cycles of PCR from streptavidin beads in multiple parallel reactions, then pooled and purified using SPRI beads.

Paired-end next generation sequencing. Two biological replicates for each of the cell lines were prepared for each target capture. Sequencing was performed on Illumina HiSeq 2500 generating 75 bp paired-end reads (Genomic Technologies Core Facility in the Faculty of Life Sciences, the University of Manchester). CASAVA software (v1.8.2, Illumina) was used to make base calls; reads failing Illumina filters were removed before further analysis. Promoter Capture libraries were each sequenced on one HiSeq lane and each Region Capture was sequenced on 0.5 of a HiSeq lane. Sequences were output in FASTQ format, poor quality reads truncated or removed as necessary, using Trimmomatic version 0.30 (ref. 25), and subsequently mapped to the human reference genome (GRCh37/hg19) and filtered to remove experimental artefacts using the Hi-C User Pipeline (HiCUP, http://www.bioinformatics.babraham.ac.uk/projects/hicup/). Off-target di-tags, where neither end mapped to a targeted HindIII restriction fragment, were removed from the final data sets using a combination of BEDTools and command line tools. Full details of the number and proportion of excluded di-tags are given in Supplementary Table 3.

Analysis of Hi-C interaction peaks. Di-tags separated by <20 kb were removed prior to analysis, as 3C data have shown a very high-interaction frequency within this distance[26]. Di-tags were then assigned to one of the four categories of ligations defined in the study by Dryden *et al.*[6] using custom scripts: (1) single baited, *cis* interaction (<5 Mb); (2) single baited *cis* interactions (>5 Mb); (3) double-baited *cis* and (4) *trans* (either single or double baited). Significant interactions for *cis* interactions within 5 Mb were determined using the 'High resolution analysis of the *cis* interaction peaks' method described in the study by Dryden *et al*[6]. To correct for experimental biases, the interactability of each fragment was determined. Interactability is calculated from the interactions from a particular baited HindIII restriction fragment to long-range, '*trans*' fragments, under the assumption that those represent random, background interactions and so should be similar in any particular baited fragment. The resulting distribution is bimodal consisting of stochastic noise (low *trans* counts) and genuine signal (high *trans* counts). A truncated negative binomial distribution was fitted to the distribution with the negative binomial truncation point for interacting restriction fragments set at a count of 3,000 and non-interacting set at 1,500 for the Promoter Capture and 600 for the Region Capture due to differences in read depth. The 5% quantile point of the non-truncated distribution was determined to provide the noise threshold. For both cell lines in both captures, the noise threshold was determined to be 400 di-tags and therefore all restriction fragments with fewer than 400 di-tags were filtered out. A negative binomial regression model was fitted to the filtered data correcting for the interactability of the captured restriction fragment and interaction distance. For interactions, where both the target and baited region were captured (double-baited interactions), we also accounted for the interactability of the other end.

We wanted to examine whether concordance between interactions called in the Region and Promoter Capture experiments increased with decreasing FDR thresholds. This is complicated because we can only define the set of interactions that could have been observed in both experiments conditional on those that were observed at a given FDR threshold in one experiment. We therefore decided to normalize to those interactions called at an FDR threshold of 20% in the region experiment and defined the following enrichment parameter: $X[i,j] = P$ (called in Region Capture at FDR i and in Promoter Capture at FDR j| called in Region Capture at FDR 20%)/P(called in Region Capture at FDR i| called in Region Capture at FDR 20%).

Interactions were considered statistically significant after combining replicates and filtering on FDR$\leq$5%. Significant Interactions were visualized in the WashU Epigenome Browser (http://epigenomegateway.wustl.edu/browser/)[27,28].

References

1. Ward, L. D. & Kellis, M. Interpreting noncoding genetic variation in complex traits and human disease. *Nat. Biotechnol.* **30,** 1095–1106 (2012).
2. Fairfax, B. P. *et al.* Innate immune activity conditions the effect of regulatory variants upon monocyte gene expression. *Science* **343,** 1246949 (2014).
3. Farh, K. K. *et al.* Genetic and epigenetic fine mapping of causal autoimmune disease variants. *Nature* **518,** 337–343 (2015).
4. Ye, C. J. *et al.* Intersection of population variation and autoimmunity genetics in human T cell activation. *Science* **345,** 1254665 (2014).
5. Schoenfelder, S., Clay, I. & Fraser, P. The transcriptional interactome: gene expression in 3D. *Curr. Opin. Genet. Dev.* **20,** 127–133 (2010).
6. Dryden, N. H. *et al.* Unbiased analysis of potential targets of breast cancer susceptibility loci by Capture Hi-C. *Genome Res.* **24,** 1854–1868 (2014).
7. Jager, R. *et al.* Capture Hi-C identifies the chromatin interactome of colorectal cancer risk loci. *Nat. Commun.* **6,** 6178 (2015).
8. Mifsud, B. *et al.* Mapping long-range promoter contacts in human cells with high-resolution capture Hi-C. *Nat. Genet.* **47,** 598–606 (2015).
9. Schoenfelder, S. *et al.* The pluripotent regulatory circuitry connecting promoters to their long-range interacting elements. *Genome Res.* **25,** 582–597 (2015).
10. Eyre, S. *et al.* High-density genetic mapping identifies new susceptibility loci for rheumatoid arthritis. *Nat. Genet.* **44,** 1336–1340 (2012).
11. Okada, Y. *et al.* Genetics of rheumatoid arthritis contributes to biology and drug discovery. *Nature* **506,** 376–381 (2014).
12. Bowes, J. *et al.* Dense genotyping of immune-related susceptibility loci reveals new insights into the genetics of psoriatic arthritis. *Nat. Commun.* **6,** 6046 (2015).
13. Hinks, A. *et al.* Dense genotyping of immune-related disease regions identifies 14 new susceptibility loci for juvenile idiopathic arthritis. *Nat. Genet.* **45,** 664–669 (2013).
14. Onengut-Gumuscu, S. *et al.* Fine mapping of type 1 diabetes susceptibility loci and evidence for colocalization of causal variants with lymphoid gene enhancers. *Nat. Genet.* **47,** 381–386 (2015).
15. Belton, J. M. *et al.* Hi-C: a comprehensive technique to capture the conformation of genomes. *Methods* **58,** 268–276 (2012).
16. Hughes, J. R. *et al.* Analysis of hundreds of *cis*-regulatory landscapes at high resolution in a single, high-throughput experiment. *Nat. Genet.* **46,** 205–212 (2014).
17. Sanyal, A., Lajoie, B. R., Jain, G. & Dekker, J. The long-range interaction landscape of gene promoters. *Nature* **489,** 109–113 (2012).
18. Grabiec, A. M. *et al.* JNK-dependent downregulation of FoxO1 is required to promote the survival of fibroblast-like synoviocytes in rheumatoid arthritis. *Ann. Rheum. Dis.* **74,** 1763–1771 (2014).
19. Nakano, K., Whitaker, J. W., Boyle, D. L., Wang, W. & Firestein, G. S. DNA methylome signature in rheumatoid arthritis. *Ann. Rheum. Dis.* **72,** 110–117 (2013).
20. Davison, L. J. *et al.* Long-range DNA looping and gene expression analyses identify DEXI as an autoimmune disease candidate gene. *Hum. Mol. Genet.* **21,** 322–333 (2012).
21. Rao, S. S. P. *et al.* A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell* **159,** 1665–1680 (2014).
22. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26,** 841–842 (2010).
23. Schneider, U., Schwenk, H. U. & Bornkamm, G. Characterization of EBV-genome negative 'null' and 'T' cell lines derived from children with acute lymphoblastic leukemia and leukemic transformed non-Hodgkin lymphoma. *Int. J. Cancer* **19,** 621–626 (1977).
24. van Berkum, N. L. *et al.* Hi-C: a method to study the three-dimensional architecture of genomes. *J. Vis. Exp.* **39,** 1869 (2010).
25. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30,** 2114–2120 (2014).
26. Naumova, N., Smith, E. M., Zhan, Y. & Dekker, J. Analysis of long-range chromatin interactions using Chromosome Conformation Capture. *Methods* **58,** 192–203 (2012).
27. Zhou, X. *et al.* The human epigenome browser at Washington University. *Nat. Methods* **8,** 989–990 (2011).
28. Zhou, X. *et al.* Exploring long-range genome interactions using the WashU Epigenome Browser. *Nat. Methods* **10,** 375–376 (2013).

Acknowledgements

We thank Frank Dudbridge for providing the R scripts to analyse the interaction data. We would like to acknowledge the Faculty of Life Sciences Genomics Facility, the assistance given by IT Services and the use of the Computational Shared Facility at The University of Manchester. This work was funded by Arthritis Research UK (grant numbers 20385, 20571 (K.D.)); Wellcome Trust Research Career Development Fellowship (G.O., AM 095684); Wellcome Trust (097820/Z/11/B); S.E. is supported through the European Union's FP7 Health Programme, under the grant agreement FP7-HEALTH-F2-2012-305549 (Euro-TEAM). A.Y. is supported by the Innovative Medicines Initiative (BeTheCure project 115142); C.W. by the Wellcome Trust (089989); C.W. and N.C. by the Wellcome Trust (091157), JDRF (9-2011-253) and the National Institute for Health Research (NIHR) Cambridge Biomedical Research Centre. The Cambridge Institute for Medical Research (CIMR) is in receipt of a Wellcome Trust Strategic Award (100140). The research leading to these results has received funding from the European Union's 7th Framework Programme (FP7/2007-2013) under grant agreement no.241447 (NAIMIT) and supported by the National Institute for Health Research Manchester Musculoskeletal Biomedical Research Unit (S.E., A.B., J.W.). P.F. and S.S. were supported by Biotechnology and Biological Sciences Research Council UK grant BBS/E/B/000C0405. The views expressed in this publication are those of the authors and not necessarily those of the NHS, the National Institute for Health Research or the Department of Health.

Author contributions

P.M., G.O., S.E. and P.F. contributed with conception and experimental design. A.M., P.M., N.C. and C.W. helped with acquisition of data. P.M., A.M., G.O., K.D., A.Y., C.W. and S.E. carried out analysis and interpretation of data. S.S., A.B., J.W., N.C. and C.W. provided administrative, technical or material support. S.E., P.M., A.M., G.O., K.D., C.W., A.B., P.F. and J.W. wrote the manuscript. S.E. supervised the study.

Additional information

Accession codes: Raw data and HindIII restriction fragment interaction counts are available in the NCBI Gene Expression Omnibus (GEO; http://www.ncbi.nlm.nih.gov/geo/) under accession number GSE69600.

Supplementary Information accompanies this paper at http://www.nature.com/naturecommunications

Competing financial interests: The authors declare no competing financial interests.

Reprints and permission information is available online at http://npg.nature.com/reprintsandpermissions/

How to cite this article: Martin, P. *et al.* Capture Hi-C reveals novel candidate genes and complex long-range interactions with related autoimmune risk loci. *Nat. Commun.* 6:10069 doi: 10.1038/ncomms10069 (2015).

Publication 5: Capture Hi-C identifies a novel causal gene, *IL20RA*, in the pan-autoimmune genetic susceptibility region 6q23

McGovern *et al. Genome Biology* (2016) 17:212
DOI 10.1186/s13059-016-1078-x

Genome Biology

RESEARCH Open Access

Capture Hi-C identifies a novel causal gene, *IL20RA*, in the pan-autoimmune genetic susceptibility region 6q23

Amanda McGovern[1], Stefan Schoenfelder[2], Paul Martin[1], Jonathan Massey[1], Kate Duffus[1], Darren Plant[1,3], Annie Yarwood[1], Arthur G. Pratt[4], Amy E. Anderson[4], John D. Isaacs[4], Julie Diboll[4], Nishanthi Thalayasingam[4], Caroline Ospelt[5], Anne Barton[1,3], Jane Worthington[1,3], Peter Fraser[2], Stephen Eyre[1] and Gisela Orozco[1*]

Abstract

Background: The identification of causal genes from genome-wide association studies (GWAS) is the next important step for the translation of genetic findings into biologically meaningful mechanisms of disease and potential therapeutic targets. Using novel chromatin interaction detection techniques and allele specific assays in T and B cell lines, we provide compelling evidence that redefines causal genes at the 6q23 locus, one of the most important loci that confers autoimmunity risk.

Results: Although the function of disease-associated non-coding single nucleotide polymorphisms (SNPs) at 6q23 is unknown, the association is generally assigned to *TNFAIP3*, the closest gene. However, the DNA fragment containing the associated SNPs interacts through chromatin looping not only with *TNFAIP3*, but also with *IL20RA*, located 680 kb upstream. The risk allele of the most likely causal SNP, rs6927172, is correlated with both a higher frequency of interactions and increased expression of *IL20RA*, along with a stronger binding of both the NFκB transcription factor and chromatin marks characteristic of active enhancers in T-cells.

Conclusions: Our results highlight the importance of gene assignment for translating GWAS findings into biologically meaningful mechanisms of disease and potential therapeutic targets; indeed, monoclonal antibody therapy targeting IL-20 is effective in the treatment of rheumatoid arthritis and psoriasis, both with strong GWAS associations to this region.

Keywords: Autoimmunity, Single nucleotide polymorphisms (SNP), Genome-wide association studies (GWAS), Causal genes, Functional genomics, Capture Hi-C

Background

In recent years, understanding of the genetic predisposition to human complex diseases has been dramatically enhanced through the application of well-powered genome-wide association studies (GWAS). Thousands of genetic variants (single nucleotide polymorphisms or SNPs) have been associated with disease [1], but the functional role of the vast majority of these disease variants is yet to be explored. This is due to the fact that around 90 % lie outside known coding regions of the genome and, therefore, their potential role in pathological mechanisms is not obvious [2, 3]. There is now strong evidence supporting a role for these non-coding variants in transcriptional regulation as they are enriched in cell type and stimulus-specific enhancer regions [4–6], which are capable of influencing their target genes through long-range chromosomal interactions [7–10]. Traditionally, GWAS associated variants have been annotated with the closest or most biologically relevant candidate gene within arbitrarily defined distances. However, this approach has been challenged by recent chromatin looping interaction studies showing

* Correspondence: gisela.orozco@manchester.ac.uk
[1] Arthritis Research UK Centre for Genetics and Genomics, Division of Musculoskeletal and Dermatological Sciences, School of Biological Sciences, Faculty of Biology, Medicine and Health, Manchester Academic Health Science Centre, The University of Manchester, Stopford Building, Oxford Road, Manchester M13 9PT, UK
Full list of author information is available at the end of the article

that interactions between enhancers and their target genes can occur over unexpectedly large genetic distances, often bypassing the nearest genes [11–13].

In order to link GWAS associated variants with disease-causing genes, we have employed a hypothesis-free method that enables the targeted characterisation of chromatin interactions at the genome-wide level at high resolution. While chromosome conformation capture studies utilising chromosome conformation capture (3C), chromosome conformation capture-on-chip (4C) and chromosome conformation capture carbon copy (5C) have been successfully used to identify interactions between regulatory elements and target genes [14–16], regions of interest and potential targets have to be considered a priori. By contrast, Hi-C allows interrogation of all interactions on a genome-wide scale [17], but the approach lacks resolution. Recently, a new method that incorporates a targeted sequence capture step into Hi-C, Capture Hi-C (CHi-C), has been developed [13, 18–20]. The method has facilitated the identification of interactions between non-coding SNPs associated with cancer and autoimmunity with their targets [18, 19, 21].

The chromosomal region 6q23 contains several variants associated with many autoimmune diseases. These associations have been annotated to the *TNFAIP3* gene, the closest most plausible causal gene within the locus, with independent variants within the gene associated with different diseases. There are three distinct linkage disequilibrium (LD) blocks independently associated with a range of autoimmune diseases, including rheumatoid arthritis (RA), systemic lupus erythematosus (SLE), celiac disease (CeD), type 1 diabetes (T1D), inflammatory bowel disease (IBD), psoriasis (Ps) and psoriatic arthritis (PsA) [22–29]. One region, containing SNPs associated with RA, SLE, CeD, IBD and T1D, tagged by the rs6920220 SNP, lies a considerable distance (>181 kb) from the *TNFAIP3* gene and its functional role has, so far, been underexplored (Fig. 1g). The second, independent association signal, tagged by rs7752903, and predisposing to RA, SLE and CeD, spans around 100 kb and includes the *TNFAIP3* gene (Fig. 1h). There is evidence that a TT > A polymorphism located within this LD block, 42 kb downstream of *TNFAIP3*, alters A20 (the protein encoded by *TNFAIP3*) expression through impaired delivery of NFκB to the *TNFAIP3* promoter [9, 30, 31]. An additional association signal, tagged by rs610604, confers risk to Ps and PsA (Fig. 1i).

The aim of the current work was to identify causal disease genes and refine the likely causal SNPs at the autoimmunity locus 6q23 by studying long-range chromatin interactions using CHi-C, to validate findings using genotype specific 3C and augment the evidence further with cell-type and genotype specific expression quantitative trait loci (eQTL) and chromatin immunoprecipitation (ChIP) analysis. Here, we report a new causal candidate disease gene within the 6q23 region, *IL20RA*, which encodes one of the subunits of the receptor for the pro-inflammatory cytokine IL-20. Our results suggest that non-coding SNPs associated with RA, SLE, CeD, IBD and T1D alter a regulatory element of *IL20RA*, some 680 kb away, which acts through long-range interactions with the *IL20RA* promoter, resulting in increased expression of the gene.

Results

6q23 variants interact with several genes, including *IL20RA*, through chromatin looping

Investigation of chromatin interactions at the 6q23 locus was carried out as part of a larger study that included all known risk loci for RA, JIA, PsA and T1D [21]. We selected four target regions mapping to 6q23 for enrichment in two different CHi-C experiments: first, the Region Capture Hi-C targeted the LD blocks ($r^2 > 0.8$) for three SNPs associated with the diseases under study: rs6920220 (RA, T1D, JIA), rs7752903 (RA) and rs610604 (Ps, PsA) (Fig. 1a–c); second, the Promoter Capture targeted all known gene promoters overlapping the region 500 kb upstream and downstream of the lead disease associated SNPs (Fig. 1d and e). CHi-C libraries were generated for two cell lines: GM12878, a B-lymphoblastoid cell line, and Jurkat, a CD4+ T-lymphoblastoid cell line.

The LD block containing the intergenic 6q23 SNP, rs6920220, targeted in the region capture, spans 47.3 kb (chr6:137959235–138006504) and contains seven restriction fragments (Fig. 1b, c and g). Of these, five were involved in statistically significant interactions. This intergenic region, containing SNPs associated with multiple autoimmune diseases, demonstrated a complex pattern of interactions, shown in Fig. 1k–n. Intriguingly, these long-range interactions involved robust and compelling interactions with both *IL20RA* and *IFNGR1*, reflecting putative roles in regulating the expression of these genes. There is also evidence of interactions with the long non-coding RNAs (lncRNAs) RP11-10J5.1 and RP11-240M16.1 downstream of the *TNFAIP3* gene.

The Region Capture experiments targeting both the LD block containing RA (rs7752903) and Ps/PsA (rs610604) associated variants, and spanning the *TNFAIP3* gene along with its upstream and downstream regions (Fig. 1h and i), showed interactions with a region proximal to the rs6920220 LD block, encompassing the lncRNAs RP11-95M15.2 (a *PTPN11* pseudogene) and RP11-356I2.1, the miRNA AL357060.1 and also an upstream region containing non-coding RNAs (Y_RNA and RP11-356I2.2) (Fig. 1k). Finally, the Region Capture experiment detected an interaction involving *TNFAIP3* and a region containing the lncRNAs RP11-10J5.1 and RP11-240M16.1 approximately 50 kb downstream of the gene, which in turn, also interacts with the intergenic rs6920220-tagged LD

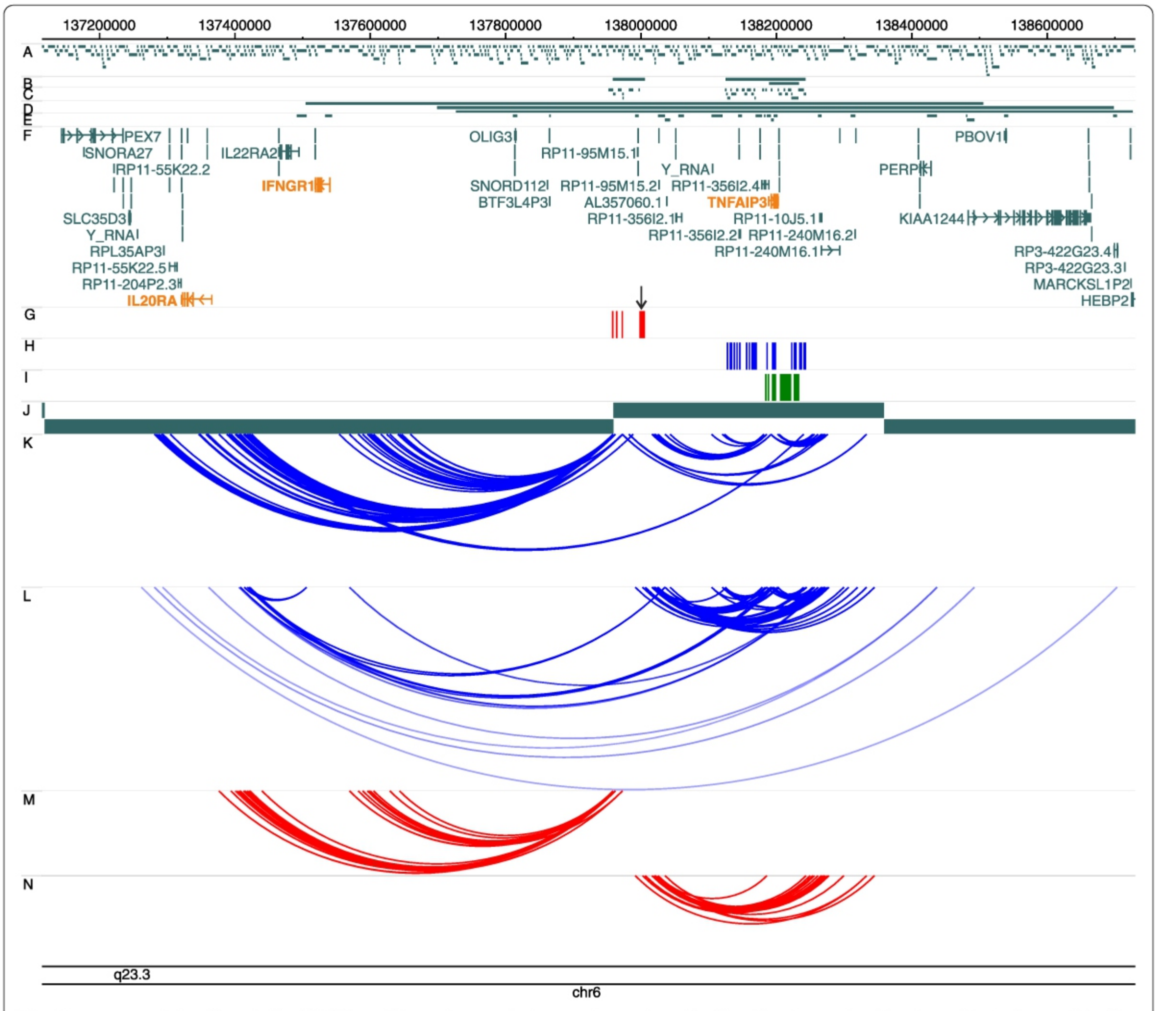


Fig. 1 Long-range interactions in the 6q23 locus. Genomic co-ordinates are shown along the *top* of each *panel* and *tracks* are labelled **a**–**n**. **a** HindIII restriction fragments. **b**–**e** Regions targeted and restriction fragments included in the Region (**b**, **c**) and Promoter (**d**, **e**) Capture experiments. **f** GENCODE V17 genes. **g**–**i** 1000 Genomes SNPs in LD (r2 ≥ 0.8) with the index SNPs rs6920220, associated with RA, SLE, celiac disease, T1D and IBD (**g**), rs7752903, associated with RA, SLE and celiac disease (**h**) and rs610604, associated with Ps and PsA (**i**). **j** Topologically associated domains (TADs) in GM12878 cells [20]. **k**–**n** Significant interactions identified in the Region and Promoter capture experiments in GM12878 (**k**, **l**) and Jurkat (**m**, **n**) cells. The *black arrow* indicates the position of the rs6927172 SNP

block. Interestingly, this region, downstream of *TNFAIP3*, showed an additional long-range interaction with the *IL20RA* gene (Fig. 1k).

These interactions were independently validated in the second, separate Promoter Capture experiment (Fig. 1d, e, l and n). Furthermore, we detected an interaction between the promoters of *TNFAIP3* and *IL20RA* that was not revealed in the Region Capture experiment, as promoters were excluded from the Region Capture experiment (Fig. 1l).

Importantly, we sought validation of CHi-C results by 3C-quantitative real-time polymerase chain reaction (qPCR). Higher interaction frequencies were confirmed for all interrogated regions, compared to adjacent non-interacting regions (Fig. 2).

To validate our analysis method, we reanalysed our CHi-C data using a recently developed analytical algorithm, CHiCAGO (Capture HiC Analysis of Genomic Organisation (http://biorxiv.org/content/early/2015/10/05/028068). The pattern of chromatin loops obtained when we applied CHiCAGO was more complex, although it confirmed our findings (Additional file 1: Figure S1). Additional interactions not passing the significance threshold in the initial analysis were found between *IL22RA2* and

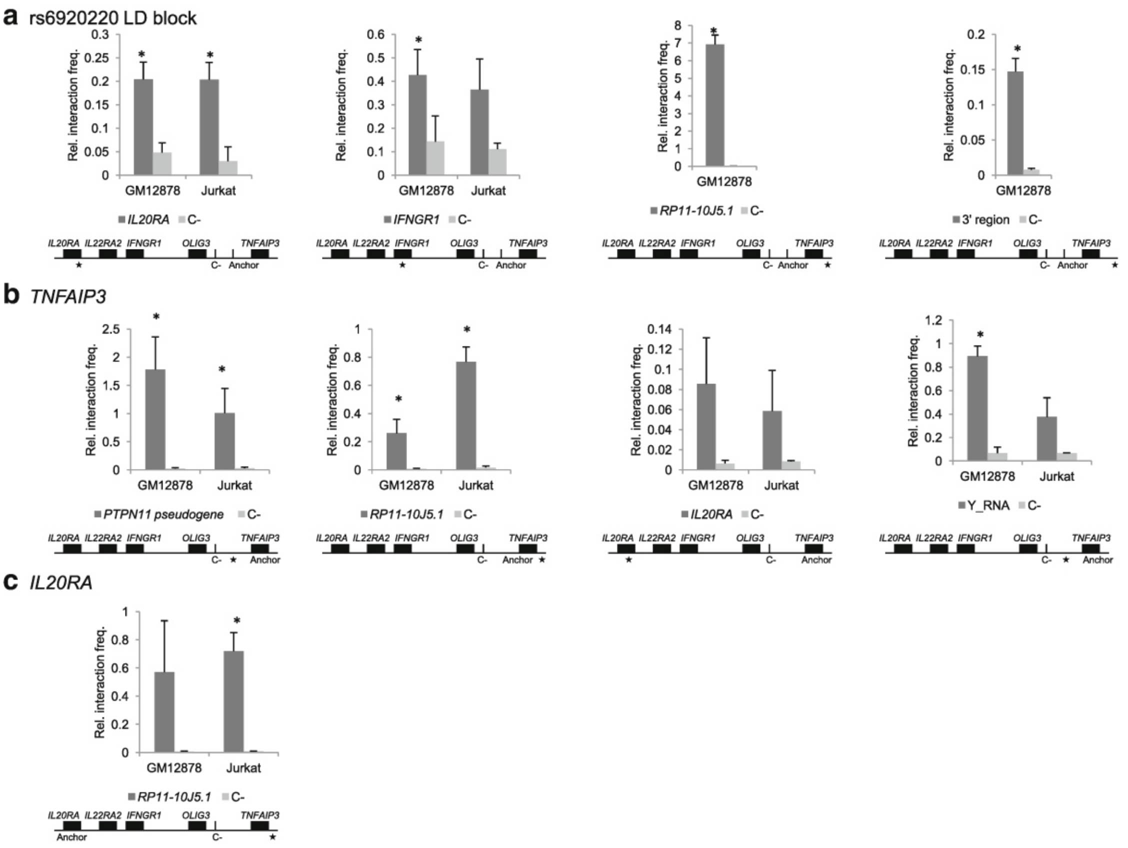


Fig. 2 Validation of CHi-C results by 3C-qPCR in GM12878 and Jurkat cell lines. The *graphs* show the relative interaction frequency of (**a**) the 6q23 intergenic disease SNPs tagged by rs6920220, (**b**) the *TNFAIP3* gene and (**c**) the *IL20RA* gene with their respective targets (*dark grey*) compared to control, non-interacting fragments (C-, *light grey*). *Diagrams* below each *graph* show the approximate location of the primers for the anchor, negative control (C-) and target (★) regions. *Error bars* indicate standard deviation of three biological replicates; * indicates t-test *P* value <0.05

the rs6920220 LD block, *IL22RA2* and the RP11-10J5.1 and RP11-240M16.1 lncRNAs downstream of *TNFAIP3*, *IFNGR1* and the rs6920220 LD block and *IFNGR1* and *TNFAIP3*. Further investigations will be required to validate these interactions.

Therefore, using CHi-C and validated by 3C-qPCR, we have confirmed that an intergenic region containing SNPs associated with RA, T1D, SLE, CeD and IBD, tagged by rs6920220 interacts with *IL20RA*, *IFNGR1* and the lncRNAs RP11-10J5.1 and RP11-240M16.1. We also confirmed that a second region, containing *TNFAIP3* and SNPs associated with RA, SLE, CeD, PsA and Ps, interacts with *IL20RA*, and a number of lncRNAs, including RP11-10J5.1 and RP11-240M16.1.

rs6927172 shows the most regulatory potential among all SNPs in LD with the top GWAS signal

Although rs6920220 is associated with a host of autoimmune diseases, its intergenic location and underexplored functional role means no causal gene has so far been unequivocally assigned. We therefore focused our work on this SNP region. The autoimmunity associated SNP rs6920220 is in tight LD ($r^2 > 0.8$) with eight other SNPs (rs6933404, rs62432712, rs2327832, rs928722, rs6927172, rs35926684, rs17264332 and rs11757201). After confirmation that these SNPs are involved in long-range interactions with several genes, including *IL20RA*, *IFNGR1*, and several lncRNAs, we aimed to narrow down the most plausible causal SNP using bioinformatics. Haploreg v4.1 was used to identify SNPs with regulatory potential [32], showing that rs6927172 demonstrates a number of lines of evidence to support a function in disease causality, including mapping to an enhancer in B-lymphoblastoid cell lines, primary stimulated Th17, and T-regulatory cells (ChromHMM chromatin state). It also maps to a region of open chromatin, characterised by DNase hypersensitivity, shows evidence of binding regulatory proteins and lies in a conserved region (Table 1). Furthermore, analysis of a

Table 1 Functional annotation of SNPs in the 6q23 intergenic LD block tagged by rs6920220 using Haploregv4.1

pos (hg19)	LD (r^2)	Variant	Ref	Alt	AFR freq	AMR freq	ASN freq	EUR freq	SiPhy cons	Promoter histone marks	Enhancer histone marks	DNAse	Proteins bound	Motifs changed
chr6:137959235	0.89	rs6933404	T	C	0.11	0.11	0.00	0.17			Blood[a]	GI[b]		STAT
chr6:137964697	0.88	rs62432712	A	G	0.08	0.10	0.00	0.17						Pax7,RORalpha1,Vax2
chr6:137973068	0.93	rs2327832	A	G	0.11	0.11	0.00	0.17			5 tissues[c]	GI,GI,PLCNT[d]		10 altered motifs[e]
chr6:137973832	0.92	rs928722	C	T	0.11	0.11	0.00	0.17			GI, PLCNT, LIV[f]			4 altered motifs[g]
chr6:137999562	0.84	rs35926684	GA	G	0.14	0.12	0.00	0.18			BLD[h]			4 altered motifs[i]
chr6:138002175	1	rs6927172	C	G	0.12	0.10	0.00	0.17	Yes	4 tissues[j]	7 tissues[k]	14 tissues[l]	13 bound proteins[m]	8 altered motifs[n]
chr6:138003822	1	rs11757201	G	C	0.08	0.10	0.00	0.17						Mrg,Sp4
chr6:138005515	1	rs17264332	A	G	0.12	0.10	0.00	0.17			BLD[o]			5 altered motifs[p]
chr6:138006504	1	rs6920220	G	A	0.12	0.10	0.00	0.17						Hltf

The following settings were used: LD threshold, $r^2 \geq 0.8$; 1000G Phase 1 population for LD calculation: EUR; Source for epigenomes: ChromHMM (25-state model using 12 imputed marks); Mammalian conservation algorithm: SiPhy-omega

[a] Dnd41 TCellLeukemia cell line

[b] Small intestine

[c] Fetal intestine small, fetal intestine large, rectal mucosa donor 31, duodenum mucosa, small intestine, rectal mucosa donor 29, stomach mucosa, placenta, fetal muscle leg, duodenum smooth muscle, fetal stomach, sigmoid colon, colonic mucosa, fetal adrenal gland, HepG2 hepatocellular carcinoma cell line

[d] Fetal intestine large, fetal intestine small, placenta

[e] BCL, GR, HDAC2, Irf, Nanog, Pou1f1, Pou2f2, RXRA, STAT, P300

[f] Duodenum mucosa, fetal intestine large, fetal intestine small, placenta, rectal mucosa donor 31, small intestine, stomach mucosa, HepG2 hepatocellular carcinoma cell line

[g] BCL, NRSF, Smad, Whn

[h] Primary monocytes from peripheral blood, primary neutrophils from peripheral blood, primary B cells from cord blood, primary T helper cells PMA-I stimulated, primary T helper 17 cells PMA-I stimulated, GM12878 lymphoblastoid cells, monocytes-CD14+ RO01746 cells

[i] CIZ, Foxd3, HDAC2, Nanog

[j] Primary T helper naïve cells from peripheral blood, primary B cells from peripheral blood, primary natural killer cells from peripheral blood, primary hematopoietic stem cells G-CSF-mobilised female, primary hematopoietic stem cells short-term culture, adipose nuclei, duodenum smooth muscle, colon smooth muscle, rectal mucosa donor 29, stomach mucosa, duodenum mucosa, liver

[k] A549 EtOH 0.02 pct lung carcinoma cell line, HeLa-S3 cervical carcinoma cell line, primary mononuclear cells from peripheral blood, primary T cells effector/memory enriched from peripheral blood, primary T cells from cord blood, primary T regulatory cells from peripheral blood, primary T helper cells from peripheral blood, primary T helper cells PMA-I stimulated, primary T helper 17 cells PMA-I stimulated, primary T helper memory cells from peripheral blood 1, primary T helper memory cells from peripheral blood 2, primary T CD8+ memory cells from peripheral blood, primary T helper naïve cells from peripheral blood, primary T CD8+ naïve cells from peripheral blood, primary monocytes from peripheral blood, primary B cells from cord blood, primary hematopoietic stem cells, primary hematopoietic stem cells G-CSF-mobilised male, primary neutrophils from peripheral blood, bone marrow derived cultured mesenchymal stem cells, Dnd41 TCellLeukemia cell line, GM12878 lymphoblastoid cells, HUVEC umbilical vein endothelial primary cells, monocytes-CD14+ RO01746 primary cells, osteoblast primary cells, mesenchymal stem cell derived adipocyte cultured cells

[l] A549 EtOH 0.02 pct lung carcinoma cell line, HeLa-S3 cervical carcinoma cell line, primary monocytes from peripheral blood, GM12878 lymphoblastoid cells, HUVEC umbilical vein endothelial primary cells, monocytes-CD14+ RO01746 primary cells, foreskin fibroblast primary cells skin01, HSMM cell derived skeletal muscle myotubes cells, primary hematopoietic stem cells G-CSF-mobilised female, primary B cells from peripheral blood, H1 derived mesenchymal stem cells, foreskin fibroblast primary cells skin02, foreskin keratinocyte primary cells skin02, HSMM skeletal muscle myoblasts cells

[m] GR (A549), ERALPHA_A (ECC-1), IRF4 (GM12878), CFOS (HUVEC), CJUN (HUVEC), GATA2 (HUVEC), CEBPB (HeLa-S3), JUND (HeLa-S3), P300 (HeLa-S3), STAT1 (HeLa-S3), MAFK (K562), STAT3 (MCF10A-Er-Src), KAP1 (U2OS)

[n] BCL, ERalpha-a, Ets, LXR, NFkB, RORalpha1, RXRA, STAT

[o] Dnd41 TCellLeukemia cell line

[p] Foxp1, HDAC2, Hoxa10, Hoxa9, Hoxd10

library of transcription factor binding site position weight matrices predicts that the SNP alters the binding site of eight transcription factors, including NFκB and BCL3 [32]. Additionally, investigation of functional annotation using RegulomeDBVersion 1.1 assigned the highest score to rs6927172 [33] (Additional file 1: Table S1). This evidence suggests that rs6927172 shows the most regulatory potential of those in LD with rs6920220. In support of this, a previous study showed evidence of differential transcription factor binding to rs6927172 alleles [34].

The risk allele of the intergenic 6q23 variant rs6927172 correlates with increased expression of *IL20RA*

We next focused on confirming disease causal genes by exploring the effect of SNP genotype on gene expression levels. However, publicly available eQTL data from different human tissues, including B-lymphoblastoid cell lines (LCLs), revealed no cis-eQTLs with the disease-associated SNPs (rs6920220, rs7752903 and rs610604) or SNPs in LD ($r^2 > 0.8$) with them.

Since gene expression is cell type specific, the effect of SNPs on transcription may occur in disease-relevant cell types only. To study the correlation between 6q23 SNP genotypes and gene expression levels in autoimmune relevant cell types, whole genome expression data from CD4+ and CD8+ primary T-cells obtained from 21 individuals from the Arthritis Research UK National Repository of Healthy Volunteers (NRHV) were interrogated. In CD4+ T-cells, the risk allele of rs6927172 correlated with increased expression of the *IL20RA* gene (Fig. 3a, *P* = 0.02), supporting that the physical interaction between them plays a functional role in the transcriptional control of *IL20RA* (Fig. 1). Additionally, CD4+ T-cell whole genome expression data were available from a cohort of 102 early undifferentiated arthritis patients collected at baseline. To avoid confounding by clinical epiphenomena typically seen in patients, individuals that were diagnosed with RA after follow-up were not included in the analysis. The correlation between rs6927172 risk alleles and increased expression of *IL20RA* was validated in this larger cohort (Fig. 3b, *P* = 0.03). No correlation was found between disease-associated SNPs (rs6927172, rs7752903 or rs610604) and expression of the previously assumed target, *TNFAIP3*, or the other interacting genes, including *IFNGR1*, in any of the CD4+ or CD8+ T-cell cohorts. Whole genome expression data were also available in primary CD19+ B-cells for the same cohort, but no eQTLs were detected for rs6927172, rs7752903 or rs610604, suggesting that the effect of rs6927172 on *IL20RA* expression may either be T-cell type specific or stimulation-dependent in B-cells. Therefore, the eQTL results showing that 6q23 non-coding variants are correlated with *IL20RA* messenger RNA (mRNA) expression in CD4+ T-cells further support that *IL20RA* is one of the target genes in the region, as evidenced by the CHi-C experiment.

rs6927172 risk allele shows higher frequency of interactions with *IL20RA* and *IFNGR1*

Having established that the non-coding 6q23 SNPs interact with several genes by long-range chromatin looping, we investigated whether the different alleles of rs6927172, the most likely candidate regulatory SNP according to bioinformatic analysis, interact with different affinities with their targets. Evaluation of 3C interactions was carried out in LCLs, as they have been genotypically well characterised

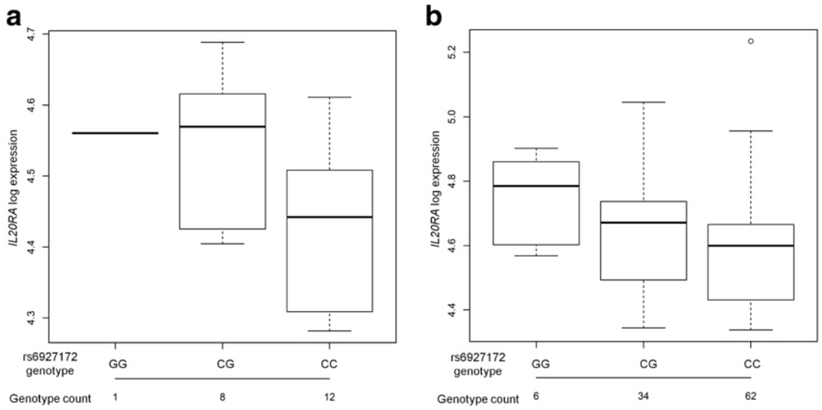


Fig. 3 eQTL effect of rs6927172 on gene expression. **a** Increased expression of *IL20RA* in primary CD4+ T-cells from 21 healthy individuals carrying the G risk allele of rs6927172, *P* = 0.02. **b** Increased expression of *IL20RA* in primary CD4+ T-cells from 102 early inflammatory arthritis clinic patients carrying the risk G allele of rs6927172, *P* = 0.03. The three different genotypes for the SNPs are displayed on the *x-axis* and gene expression levels on the *y-axis*. *Error bars* indicate standard deviation

as part of the HapMap Project and cells carrying the three different genotypes for the rs6927172 variant (GM11993 CC, GM12878 CG and GM07037 GG) are readily accessible commercially. This experiment revealed significantly higher interaction frequencies between both *IL20RA* and *IFNGR1* and the restriction fragment containing rs6927172 in individuals carrying the risk G allele of this SNP compared with the homozygous non-risk allele (GG versus CC, *P* = 0.01; CG versus CC, *P* = 0.01 and GG versus CC, *P* = 0.04; CG versus CC, *P* = 0.02, respectively) (Fig. 4). Interaction frequencies between the fragment containing rs6927172 and both fragments containing the lncRNAs RP11-10J5.1 and RP11-240M16.1 were similar regardless of genotype (Additional file 1: Figure S2). Similarly, none of the interactions between *TNFAIP3* and targets identified in the CHi-C experiment (*PTPN11* pseudogene, RP11-10J5.1, RP11-240M16.1, Y_RNA and *IL20RA*) and between *IL20RA* and RP11-10J5.1 were influenced by rs6927172 genotype (Additional file 1: Figure S3).

6q23 is one of the most important loci for RA susceptibility, being the third most strongly associated region after *HLA-DRB1* and *PTPN22*. Although T-cells are thought to be the most important cell type in RA pathogenesis, synovial fibroblasts have also been shown to play a crucial role in the perpetuation of disease [35]. Therefore, we sought to evaluate the 3D conformation of the locus in this cell type. The preferential interaction of the fragment containing rs6927172 and *IL20RA* was confirmed by 3C-qPCR in primary human synovial fibroblasts (Additional file 1: Figure S4).

Hence, our experiments suggest that increased *IL20RA* expression that correlates with the risk G allele of rs6927172 may be mediated through increased ability to bind the *IL20RA* gene via chromatin looping.

The risk allele of rs6927172 shows increased enrichment of regulatory proteins

To further explore the role of rs6927172 in transcriptional regulation, we evaluated enrichment of chromatin marks of active regulatory elements to this site using chromatin immunoprecipitation (ChIP) in LCLs. We observed an enrichment of histone marks, H3K4me1 and H3K27ac, to the region containing the SNP, compared to a non-regulatory control region (*P* = 0.0001 and *P* = 0.0001, respectively) and to a no antibody control sample (*P* = 0.0001 and *P* = 0.0008, respectively), confirming the bioinformatic evidence that rs6927172 is located in a regulatory element (Additional file 1: Figure S5). We then performed allele-specific qPCR using Taqman probes complementary to each rs6927172 allele in Jurkat T-cells and GM12145 B-cells, which are both heterozygous for the variant, and the balance between the immunoprecipitated fragments with the C allele or the G allele was determined. In Jurkat cells, the risk G allele showed evidence of increased enrichment of both H3K4me1 (*P* = 0.009) and H3K27ac (*P* = 0.03), compared to the non-risk allele (Fig. 5), supporting the CD4+ specific nature of the eQTL evidence and further suggesting that the risk allele is correlated with an increased regulatory activity. By contrast, in GM12145 B-cells, where no eQTL evidence was detected/observed, the non-risk C allele showed evidence of increased enrichment for histone marks (*P* = 0.009 and *P* = 0.0001 for H3K4me1 and H3K27ac respectively), further highlighting the cell type specificity of transcriptional regulation (Additional file 1: Figure S5).

The rs6927172 variant was predicted to alter the binding motif for eight transcription factors, including NFκB and BCL3 (Table 1). Since NFκB is an important mediator

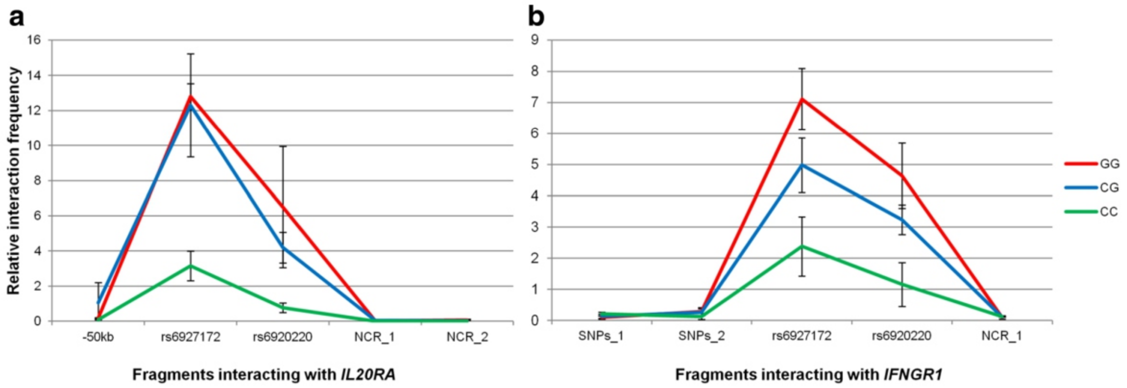


Fig. 4 Genotype-specific 3C showing preferential interaction of the disease risk G allele of rs6927172 with *IL20RA* (**a**) and *IFNGR1* (**b**). *–50 kb* restriction fragment located 50 kb upstream of the rs6927172 containing restriction fragment, *rs6927172* restriction fragment containing rs6927172, *rs6920220* restriction fragment containing the top GWAS SNP in the 6q23 region, *NCR* non-interacting control region. *Error bars* indicate standard deviation of three biological replicates

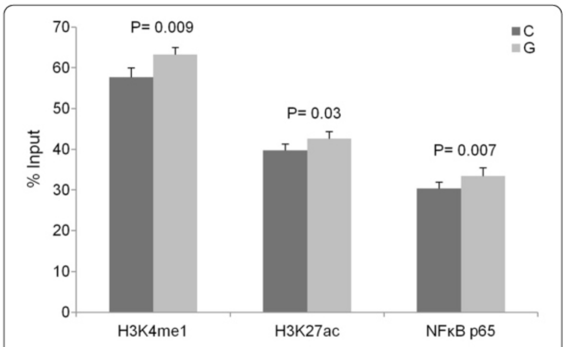


Fig. 5 Allele-specific ChIP in Jurkat cells, showing increased binding of H3K4me1, H3K27ac and NFκB p65 to the risk allele of rs6927172. *Error bars* indicate standard deviation of three biological replicates

of the immune response [36] and previous studies have shown that the TT > A variant, which maps to the *TNFAIP3* LD block tagged by rs7752903, impairs the binding of this transcription factor [9], we experimentally tested whether NFκB binds rs6927172 alleles with different affinities. We performed ChIP in Jurkat and GM12878 cell lines using antibodies for the p50 and p65 subunits of NFκB. Estimation of the C/G ratio in the immunoprecipitated chromatin was performed and results showed that, in Jurkat cells, the p65 subunit of NFκB binds with higher affinity to the risk G allele, compared to the non-risk C allele ($P = 0.007$) (Fig. 5). The SNP did not show evidence of altered binding of NFκB in the B-lymphoblastoid cell line.

BCL3 is a transcriptional co-activator that inhibits the nuclear translocation of the NFκB p50 subunit in the cytoplasm and contributes to the regulation of transcription of NF-κB target genes in the nucleus [37–39]. Therefore, we also investigated binding of BCL3 to the different alleles of rs6927172 using the same approach. Although this transcription factor seems to be part of the transcriptional machinery at the site of the SNP, BCL3 binding showed no statistically significant differences between the two alleles, either in Jurkat or in GM12878 cells.

Taken together, these results suggest that the mechanism by which the risk allele of rs6927172 increases expression of *IL20RA* may be mediated by an increased regulatory activity and augmented binding of the transcription factor NFκB.

Discussion

The chromosomal region 6q23 is an important locus in autoimmunity. It is an exemplar complex non-coding genomic region, some distance from the nearest gene, containing enhancer elements and implicated in multiple diseases by GWAS, but where independent variants associate with different conditions. To date, investigation of the functional consequences of disease-associated alleles have focused almost exclusively on the gene *TNFAIP3*. Here we present findings from a hypothesis-free, systematic approach using the recently developed CHi-C method to identify causal genes at this locus. Our experiments have revealed that the spatial organisation of the chromatin at this region is complex, bringing together several genes with key roles in the immune response, including *IL20RA*, *IFNGR1* and *TNFAIP3*, alongside regulatory elements containing SNPs associated with different autoimmune diseases. This supports the recently proposed concept of specialised transcription factories, where co-regulated genes come together to share transcription factors and regulatory elements such as enhancers [40].

Previous studies investigating the functional role of 6q23 disease variants had been restricted to the SNPs mapping to the LD block tagged by rs7752903 spanning the *TNFAIP3* gene, associated with SLE, RA and celiac disease, showing that the TT > A variant, located downstream of *TNFAIP3*, impairs that gene's expression through chromatin looping and altered NFκB binding [9, 30, 31, 40]. However, the functional impact of the remaining disease-associated SNPs at the locus, such as the intergenic rs6920220 nominally assigned to *TNFAIP3*, had remained unexplored. Our CHi-C study, supplemented by confirmatory 3C, eQTL and ChIP evidence, offers for the first time a firm indication that autoimmune-associated regions in general [21], and this region in particular, can demonstrate complex regulatory interactions with a number of plausible candidate genes, potentially functional lncRNA genes and, importantly, each other. The complexity of the interactions is magnified when considering the differences observed in cell types (here, in B and T-cell lines and synovial fibroblasts). Interestingly, the rs6927172 alleles, associated with RA, correlate with *IL20RA* expression levels in CD4+ T-cells, supporting the accumulating evidence that CD4+ T-cells are the most relevant cell type to RA [41]. Published high resolution Hi-C data were available for GM12878 B-lymphoblastoid cells and we observed numerous, strong interactions between the 6q23 intergenic SNPs and *IL20RA*, supporting our results [42]. In contrast, these interactions with the associated intergenic region were markedly decreased or non-existent in cell lines that do not express *IL20RA*, such as human umbilical vein endothelial cells (HUVEC) or chronic myeloid leukaemia (K562) cells (Additional file 1: Figure S7), supporting a cell type dependent regulatory role for the disease-associated enhancer region and *IL20RA*.

Chromatin looping and eQTL experiments strongly support *IL20RA* as a putative causal autoimmunity gene in 6q23. The *IL20RA* gene encodes the IL-20 receptor α subunit (IL-20RA), which can form a heterodimeric receptor with either IL-20RB to bind IL-19, IL-20 and

IL-24, or with IL-10RB to bind IL-26 [43]. Evidence suggests that this family of cytokines have a pro-inflammatory effect, and are essential in the activation of the epithelial innate immunity [44], with expression of *IL20RA* detected in whole blood, T-cells, B-cells and monocytes [45]. Recently, interactions of IL-20 subfamily cytokines with their receptors have been shown to be involved in the pathogenesis of RA. IL-20 and its receptors are upregulated in the synovium of RA patients [46–50] and IL-19, IL-20 and IL-22 are able to increase the proliferation of synovial cells and induce IL-6, IL-8 and CCL2 in these cells [48, 50]. In rats, experimentally induced autoimmune arthritis and collagen-induced arthritis are attenuated by IL-19 blockade [51] and administration of soluble IL-20RA [47, 51], respectively. These cytokines also have involvement in skin inflammation [52]. Overexpression of *Il20, Il22* or *Il24* in mice leads to the development of psoriasis-like skin lesions [53–55], and levels of IL-19, IL-20, IL-22 and IL-24 are increased in psoriatic skin [56–58]. Notably, SNPs mapping to the *TNFAIP3* region have been shown to be associated with Ps and PsA, but map to a different risk haplotype, tagged by rs610604, distinct to other autoimmune diseases [22, 26]. Very interestingly, two recent clinical trials have demonstrated that anti-IL-20 monoclonal antibody is effective in the treatment of RA and psoriasis [59, 60]. Furthermore, levels of IL-19, IL-20, IL-24 and IL-26 are also elevated in serum of patients with inflammatory bowel disease [61–64], which is associated with the intergenic 6q23 variants tagged by rs6920220 [25]. The evidence that SNPs associated with different autoimmune diseases interact with each other and the same genes supports a concept that regional genetic variation, regulating similar target genes, but with mechanistic and cellular differences, are risk factors for different diseases. This may also suggest that blocking the IL-20 pathway might be effective in the treatment of multiple autoimmune diseases. Indeed, a recent study has shown that selecting a therapeutic target with genetic data supporting its role could double the chance of a drug's success in clinical improvement [65].

Our CHi-C experiment suggested another potential novel causal gene in the 6q23 region, *IFNGR1*. In addition, targeted 3C experiments found that the interaction between rs6927172 and this gene is stronger when the disease risk G allele is present. *IFNGR1* encodes one of the subunits of the interferon gamma (IFN-γ) receptor. This cytokine plays an important role in autoimmunity, since it is involved in macrophage activation, enhanced MHC expression on neighbouring cells, balancing Th1/Th2 cell differentiation, and inducing the secretion of other pro-inflammatory cytokines [66]. Although it has been shown that an increased expression of *IFNGR1* in blood is associated with RA [67], we did not detect an effect of rs6927172 genotype on this gene's expression levels in CD4+ and CD8+ T cells. eQTLs, though, are context-specific [6, 68–72] and, therefore, it would be interesting to explore whether the SNP influences *IFNGR1* expression in other cell types and/or under different stimulatory conditions.

Whereas we provide evidence of additional putative causal genes in the 6q23 region, the *TNFAIP3* gene remains a strong candidate. The role of *TNFAIP3* in autoimmunity is well established. The protein encoded by *TNFAIP3*, A20, is induced by tumour necrosis factor (TNF) and inhibits NFκB activation and TNF-mediated apoptosis [73]. Mice deficient for A20 develop severe multiorgan inflammation [74] and tissue-specific deletion of A20 results in different phenotypes that resembles human autoimmune diseases such as inflammatory polyarthritis (macrophages), SLE (dendritic cells), IBD (intestinal epithelial cells) or psoriasis (keratinocyes) [73].

Bioinformatic analysis suggested that rs6927172 is the most likely causal SNP in the rs6920220 LD block. Genotype specific 3C showed increased interactions with the *IL20RA* gene when the risk G allele is present compared with the non-risk allele. By contrast, the genotype-specific interaction was not observed for the rs6920220 variant. However, although bioinformatic evidence and ChIP experiments coupled with previous evidence from electrophoretic mobility shift assays [34] point to rs6927172 as the most likely causal SNP, the resolution of this experiment is limited by the restriction enzyme used; rs6927172 is located in the same restriction fragment as rs35926684 and both SNPs are strongly correlated ($r^2 = 0.8$). Therefore, although bioinformatic evidence suggests that rs35926684 is less likely to affect binding of regulatory proteins, the possibility that it is the causal SNP, or that both SNPs contribute to transcriptional regulation, cannot be excluded.

Our study illustrates the challenges in linking associated variants to function. Associated variants can be linked to a number of genes, dependent on which enhancer they are located within and the cell type under investigation. This could explain apparent inconsistencies in findings; for example, how the risk variant of rs6927172 is associated with higher levels of active enhancer histone marks in Jurkat cells, but has the opposite effect in GM12878 cells. Indeed, up to 50 % of allele specific associations with epigenetic marks of enhancer activity (histoneQTLs) have been reported to show inconsistent direction of effects between samples, indicating the intricacies that exist in gene regulation [75]. Nonetheless, our work reinforces previous evidence that the nearest plausible biological candidate gene is not necessarily the causal gene. While *TNFAIP3* gene involvement is still implicated at the 6q23 locus, the primary causal gene may well be *IL20RA*, supported by the success of anti-IL20 therapies in RA and Ps.

It is noteworthy that the intergenic 6q23 SNP, correlated with higher frequency of interactions with *IL20RA*, higher expression of *IL20RA* and increased enrichment of histone marks of active enhancers and NFκB, is located at the boundary of two topologically associated domains (TADs) (Fig. 1g and j). TADs are genomic regions that show high levels of interaction within the region and little or no interaction with bordering regions and are thought to be conserved across different cell types and species [76, 77]. It has been shown that boundaries between TADs can separate functionally distinct regions of the genome [78]. Intriguingly, it has been suggested that eQTLs often occur around TAD boundaries and preferentially associate with genes across domains [79]. There is now evidence that disruption of TAD boundaries can cause ectopic interactions between regulatory non-coding DNA and gene promoters, resulting in pathogenic phenotypes [80]. Our CHi-C experiments show long-range interactions between *IL20RA* and targets located outside the TAD this gene is located, i.e. the intergenic disease-associated SNPs, *TNFAIP3* and several lncRNAs (Fig. 1). The cell lines used in these experiments (GM12878 and Jurkat) are both heterozygous for rs6927172 and genotype-specific 3C experiments showed that the interaction between this SNP and *IL20RA* occurs preferentially when the risk allele is present (Fig. 3). It would be interesting to explore whether this autoimmunity associated variant exerts its pathogenic effect through a disruption of the TAD boundary between *IL20RA* and potential regulatory elements that would not otherwise interact with it.

Conclusions

We provide evidence that an intergenic enhancer region on 6q23, associated with numerous autoimmune diseases and nominally assigned to *TNFAIP3* although over 200 kb from the nearest gene, makes allele-specific, regulatory contact with *IL20RA*, the target of an existing drug and located 680 kb away from the associated region. Our findings show how functional evaluation of disease risk loci can help better translate GWAS findings into biologically meaningful mechanisms of disease and validate existing therapeutic targets or suggest potential new ones.

Methods

Cell culture

B-lymphoblastoid cell lines (LCL) were obtained from the Coriell Institute for Medical Research (Additional file 1: Table S2). Cells were grown in vented 25 cm^2 cell culture flasks containing 10–20 mL of Roswell Park Memorial Institute medium (RPMI)-1640 + 2 mM L-glutamine culture medium, supplemented with 15 % fetal bovine serum (FBS). Flasks were incubated upright at 37 °C/5 % CO_2. Cultures were regularly monitored to maintain a cell density in the range of 2 × 10^5–5 × 10^5viable cells/mL. Cells were split every two days into fresh medium until they reached a maximum density of 1 × 10^6 cells/mL.

Jurkat E6.1 human leukaemic T-lymphoblast cells were obtained from LGC Standards. Cells were grown in vented 25 cm^2 cell culture flasks containing 10–20 mL of RPMI-1640 + 2 mM L-glutamine, supplemented with 10 % FBS. Flasks were incubated upright at 37 °C/5 % CO_2 and the cultures regularly monitored to maintain a cell density in the range of 3 × 10^5–9 × 10^5 viable cells/mL.

These cell lines are not listed in the in the database of commonly misidentified cell lines maintained by ICLAC, were authenticated using STR analysis and were tested for mycoplasma contamination (MycoSEQ® Mycoplasma Detection System, 4460625, Life Technologies).

Capture Hi-C

Chromatin interactions at 6q23 were scrutinised using CHi-C as part of a larger study that included all confirmed risk loci for four autoimmune diseases (RA, JIA, PsA and T1D) [21].

We tested chromatin interactions in two complementary experiments: Region Capture, which targeted regions associated with disease [22, 27, 81–83], and Promoter Capture, which provided independent validation by capturing all gene promoters within 500 kb upstream and downstream of lead disease-associated SNPs. Associated regions were defined by selecting all SNPs in LD with the lead disease-associated SNP ($r^2 \geq 0.8$; 1000 Genomes phase 1 EUR samples; May 2011). For the Promoter Capture experiment, HindIII restriction fragments were identified within 500 bp of the transcription start site of all genes mapping to the defined regions (*Ensembl* release 75; GRCh37). A control region with well characterised long-range interactions was also included, *HBA* [84]. Capture oligos (120 bp; 25–65 % GC, <3 unknown (N) bases) were designed using a custom Perl script within 400 bp but as close as possible to each end of the targeted HindIII restriction fragments.

Human T-cell (Jurkat) and B-cell (GM12878) lines were used, since they are among the most relevant cell types in autoimmune disease [5]. Hi-C libraries were generated as previously described [85]. Cells of 5–6 × 10^7 were grown to ~90 % confluence and cross-linked with 2 % formaldehyde for 10 min at room temperature. The cross-linking reaction was quenched by adding cold 1 M glycine to a final concentration of 0.125 M for 5 min at room temperature, followed by 15 min on ice. Cells were resuspended in 50 mL ice-cold lysis buffer (10 mM Tris–HCl pH 8, 10 mM NaCl, 0.2 % Igepal CA-

630, protease inhibitors) and lysed on ice for 30 min, with 2 × 10 strokes of a Dounce homogeniser. Following lysis, the nuclei were pelleted and washed with 1.25 × NEB Buffer 2 then resuspended in 1.25 × NEB Buffer 2. Hi-C libraries were digested using HindIII then prepared as described in van Berkum et al. [86] with modifications described in Dryden et al. [18]. Pre-Capture amplification was performed with eight cycles of PCR on multiple parallel reactions from Hi-C libraries immobilised on Streptavidin beads which were pooled post-PCR and SPRI bead purified. The final library was resuspended in 30 μL TLE (10 mM Tris pH8; 0.1 mM EDTA) and the quality and quantity assessed by Bioanalyzer and qPCR.

Hybridisation of Agilent SureSelect custom Promoter and Region Capture RNA bait libraries to Hi-C libraries was carried out using Agilent SureSelectXT reagents and protocols. Post-capture amplification was carried out using six cycles of PCR from streptavidin beads in multiple parallel reactions, then pooled and purified using SPRI beads.

Two biological replicates for each of the cell lines were prepared for each target capture. Sequencing was performed on Illumina HiSeq 2500 generating 75 bp paired-end reads (Genomic Technologies Core Facility in the Faculty of Life Sciences, University of Manchester). CASAVA software (v1.8.2, Illumina) was used to make base calls; reads failing Illumina filters were removed before further analysis. Promoter Capture libraries were each sequenced on one HiSeq lane and each Region Capture library was sequenced on 0.5 of a HiSeq lane. Sequences were output in FASTQ format, poor quality reads truncated or removed as necessary, using Trimmomatic version 0.30 [87], and subsequently mapped to the human reference genome (GRCh37/hg19) and filtered to remove experimental artefacts using the Hi-C User Pipeline (HiCUP, http://www.bioinformatics.babraham.ac.uk/projects/hicup/). Off-target di-tags, where neither end mapped to a targeted fragment, were removed from the final datasets.

Di-tags separated by <20 kb were removed prior to analysis, as 3C data have shown a very high interaction frequency within this distance [88]. Significant interactions for cis interactions within 5 Mb were determined using the 'High resolution analysis of cis interaction peaks' method described by Dryden et al. [18]. To correct for experimental biases, the interactability of each fragment was calculated to long-range, '*trans*' fragments, under the assumption that those represent random, background interactions and so should be similar in any particular baited fragment. The resulting distribution is bimodal consisting of stochastic noise (low *trans* counts) and genuine signal (high *trans* counts). A truncated negative binomial distribution was fitted to the distribution. The 5 % quantile point of the non-truncated distribution was determined to provide the noise threshold. A negative binomial regression model was fitted to the filtered data correcting for the interactability of the captured restriction fragment and interaction distance. For interactions where both the target and baited region were captured (double-baited interactions) we also accounted for the interactability of the other end.

Interactions were considered statistically significant after combining replicates and filtering on FDR ≤ 5 %. Significant interactions were visualised in the WashU Epigenome Browser [89, 90].

Chromosome conformation capture (3C)

Validation of interactions was carried out on biological replicate 3C libraries for each of the cell lines (GM12878 and Jurkat). Libraries were prepared using the cross-linking, digestion with HindIII and ligation steps used for the generation of Hi-C libraries [84] but without the biotin fill-in step. qPCR was carried out using Power SYBR® Master Mix (Life Technologies) according to the manufacturer's instructions using the following cycling conditions: 50 °C 2 min, 95 °C 10 min, followed by 40 cycles of 95 °C 15 s, 60 °C 1 min. qPCR was performed in triplicate using 50 ng of 3C library [88]. Standard curves for each primer set used in the qPCR were generated using tenfold serial dilutions of 3C control template libraries, prepared by digestion and random ligation of bacterial artificial chromosomes (BACs) (Life Technologies) spanning the region of interest with minimal overlap (Additional file 1: Table S3). Data were normalised to a short-range ligation product using the bait primer in combination with a primer for adjacent HindIII fragments, to control for differences in cross-linking and ligation efficiencies between different cell lines. 3C primers are shown in Additional file 1: Table S4. Statistical analysis was performed in STATA by paired t-test. P values < 0.05 were considered statistically significant. Variance between groups was similar (two-tailed F-test for equality of two variances $P > 0.05$).

Bioinformatics

To narrow down the most plausible causal SNP among all variants in LD with the lead GWAS SNP rs6920220, Haploreg v4.1 was used with the following settings: LD threshold, $r^2 \geq 0.8$; 1000G Phase 1 population for LD calculation: EUR; Source for epigenomes: ChromHMM (25-state model using 12 imputed marks); Mammalian conservation algorithm: SiPhy-omega. Additionally, RegulomeDBVersion 1.1 was used.

Expression quantitative trait loci (eQTLs) analysis

Public eQTL data were interrogated using Haploreg v4.1 [32], which examines all datasets obtained from the GTEx analysis release V6 (http://www.gtexportal.org/

static/datasets/gtex_analysis_v6/single_tissue_eqtl_data/GTEx_Analysis_V6_eQTLs.tar.gz), the GEUVADIS analysis (EUR and YRI panels, http://www.ebi.ac.uk/arrayexpress/files/E-GEUV-1/analysis_results/), the NCBI eQTL Browser (http://www.ncbi.nlm.nih.gov/projects/gap/eqtl/index.cgi, lymphoblastoid cell lines [91, 92], liver [93] and brain [94]) and eight additional studies including data obtained from tumours [95], blood [96], lung [97], heart [98], monocytes [4], bone [99], lymphoblastoid cell lines [100] and brain [101].

Four whole genome gene expression datasets were available: CD4+ and CD8+ T-cells from 21 healthy individuals of the National Repository of Healthy Volunteers (NRHV), The University of Manchester (North West Centre for Research Ethics Committee) (Additional files 2, 3, 4 and 5), and CD4+ T-cells and CD19+ B-cells from 102 early undifferentiated arthritis patients, Newcastle University (Newcastle and North Tyneside Local Research Ethics Committee) (Additional files 6, 7, 8 and 9). Informed consent was obtained from all participants. mRNA was isolated from sorted cell subsets, quality and concentration assessed using the Agilent Bioanalyzer and Nanodrop, before complementary DNA (cDNA)/complementary RNA (cRNA) conversion using Illumina TotalPrep RNA Amplification Kits. A total of 750 ng of cRNA was hybridised to HumanHT-12 v4 Expression BeadChip arrays according to the manufacturer's protocol before being scanned on the Illumina iScan system. Raw expression data were exported from Illumina GenomeStudio and analysed using the R Bioconductor package 'limma' [102]. Briefly, the neqc function was used for log2 transformation of the data, background correction and quantile normalisation using control probes. Principal component analysis was used to detect batch effects. The cDNA/cRNA conversion produced the largest batch effect in both cohorts and was corrected using ComBat (in R Bioconductor package sva) (http://bioconductor.org/packages/release/bioc/html/sva.html). Genome-wide genotype data were generated using the Illumina HumanCoreExomeBeadChip kit. Genotype data were aligned to the 1000 genomes reference strand, pre-phased using SHAPEIT2 (v2.r727 or v2.r790), before imputation using IMPUTE2 (v2.3.0 or v2.3.1) with the 1000 genome reference panel (Phase 1, December 2013 or June 2014). Imputed data were hard-called to genotypes using an INFO score cutoff of 0.8 and posterior probability of 0.9. The effect of the SNPs on gene expression was analysed using MatrixEQTL (v.2.1.0) (http://www.bios.unc.edu/research/genomic_software/Matrix_eQTL/) with an additive linear model. The errorCovariance = numeric() parameter was set to account for possible differences in variance between groups. SNPs within 4 Mb of a gene expression probe were considered to be cis-eQTL, since the majority (99 %) of interactions detected in the CHi-C experiment happened within a 4 Mb window. P values < 0.05 were considered statistically significant. The study (N = 102 early arthritis patients) had 80 % power to detect a change of 0.08 log expression at 5 % significance level.

Chromatin immunoprecipitation (ChIP)

1×10^7 cells were cross-linked with 1 % formaldehyde for 10 min at room temperature. Cells were lysed in 1 mL of ChIP lysis buffer (50 mM Tris–HCl pH8.1, 10 mM EDTA, 1 % SDS, one protease inhibitor cocktail tablet) and chromatin sheared using a Covaris S220 with the following conditions: target base pairs: 200–400 bp, duty cycle: 5 % for LCL; 10 % for Jurkat cells, peak incident power: 140 Watts, cycles per burst: 200, temperature: 4 °C, time: 20–25 min.

Each immunoprecipitation (IP) was carried out in triplicate using LCLs obtained from HapMap individuals (Additional file 1: Table S1). The negative control was a no antibody control or IgG. Antibodies were available from Abcam for NFκB p50 (ab7971), NFκBp65 (ab7970), H3K4me1 (ab8895) and H3K27ac (ab4729) and from Santa Cruz for BCL3 (sc-185X). To detect the relative enrichment of regions interacting with the target protein, qPCR of ChIP and input samples was carried out. qPCR was performed in triplicate using SYBR green, or TaqMan probes complementary to each allele of rs6927172 for allele-specific assays (Applied Biosystems, assay ID C___1575580_100), on an Applied Biosystems QuantStudio 12 K Flex qPCR instrument. Primers were designed for the target SNP region, positive control region and negative control region (Additional file 1: Table S5). Following qPCR, the % input for each sample was calculated and statistical analysis of ChIP data was carried out to determine significant differences in antibody binding to the different SNP genotypes in STATA by paired t-test. P values < 0.05 were considered statistically significant. Variance between groups was similar (two-tailed F-test for equality of two variances $P > 0.05$).

Additional files

Additional file 1: Supplementary tables and figures. (DOCX 2089 kb)

Additional file 2: Contains the cis-eQTLs with P value < 5 % for CD4+ T-cells obtained from NRHV participants. (TXT 405 kb)

Additional file 3: Contains the normalised expression values for CD4+ T-cells obtained from NRHV participants. (TXT 26 kb)

Additional file 4: Contains the cis-eQTLs with P value < 5 % for CD8+ T-cells obtained from NRHV participants. (TXT 352 kb)

Additional file 5: Contains the normalised expression values for CD8+ T-cells obtained from NRHV participants. (TXT 26 kb)

Additional file 6: Contains the cis-eQTLs with P value < 5 % for B-cells obtained from early undifferentiated arthritis patients. (TXT 985 kb)

Additional file 7: Contains the normalised expression values for B-cells obtained from early undifferentiated arthritis patients. (TXT 124 kb)

Additional file 8: Contains the cis-eQTLs with *P* value < 5 % for T-cells obtained from early undifferentiated arthritis patients. (TXT 1162 kb)

Additional file 9: Contains the normalised expression values for T-cells obtained from early undifferentiated arthritis patients. (TXT 125 kb)

Abbreviations

3C: Chromosome conformation capture; BACs: Bacterial artificial chromosomes; CeD: Celiac disease; CHi-C: Capture Hi-C; CHiCAGO: Capture HiC Analysis of Genomic Organisation; ChIP: Chromatin immunoprecipitation; eQTL: Quantitative trait loci; FBS: Fetal bovine serum; GWAS: Genome-wide association studies; HiCUP: Hi-C User Pipeline; IBD: Inflammatory bowel disease; IFN-γ: Interferon gamma; IL-20RA: IL-20 receptor α subunit; LCLs: B-lymphoblastoid cell lines; LD: Linkage disequilibrium; lncRNAs: Long non-coding RNAs; NRHV: National Repository of Healthy Volunteers; Ps: Psoriasis; PsA: Psoriatic arthritis; qPCR: Quantitative real-time PCR; RA: Rheumatoid arthritis; RPMI: Roswell Park Memorial Institute medium; SLE: Systemic lupus erythematosus; SNPs: Single nucleotide polymorphisms; T1D: Type 1 diabetes; TADs: Topologically associated domains; *TNFAIP3*: Tumour necrosis factor alpha-induced protein 3

Acknowledgements

We thank Dr Kathryn Steel for carrying out cell separation and gene expression profiling of healthy volunteers and Dr Nisha Nair for her contribution to the generation of gene expression data in the early arthritis patient cohort. The authors would also like to acknowledge the Faculty of Life Sciences Genomics Facility and the assistance given by IT Services and the use of the Computational Shared Facility at The University of Manchester.

Funding

This work was funded by the Wellcome Trust (grant no. 095684), Arthritis Research UK (grant no. 20385) and supported by the National Institute for Health Research Manchester Musculoskeletal Biomedical Research Unit. The views expressed in this publication are those of the authors and not necessarily those of the NHS, the National Institute for Health Research or the Department of Health.

Availability of data and materials

The CHi-C datasets supporting the conclusions of this article are available in the Gene Expression Omnibus repository, http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE69600. The in-house eQTL data are provided as supplemental data files (Additional files 2, 3, 4, 5, 6, 7, 8 and 9). Public eQTL data were interrogated using Haploreg [32], which examines all datasets obtained from the GTEx analysis release V6 (http://www.gtexportal.org/static/datasets/gtex_analysis_v6/single_tissue_eqtl_data/GTEx_Analysis_V6_eQTLs.tar.gz), the GEUVADIS analysis (EUR and YRI panels, http://www.ebi.ac.uk/arrayexpress/files/E-GEUV-1/analysis_results/), the NCBI eQTL Browser (http://www.ncbi.nlm.nih.gov/projects/gap/eqtl/index.cgi, lymphoblastoid cell lines [91, 92], liver [93] and brain [94]) and eight additional studies including data obtained from tumours [95], blood [96], lung [97], heart [98], monocytes [4], bone [99], lymphoblastoid cell lines [100] and brain [101].

Authors' contributions

AMG, SS, PM, JM, KD, DP, AP, AA, JI, JD, NT, CO and GO performed experiments and contributed to the writing of the paper. SS, PF, AB, JW, SE and GO designed experiments and wrote the paper. All authors read and approved the final manuscript.

Competing interests

The authors declare that they have no competing interests.

Consent for publication

Not applicable

Ethics approval and consent to participate

Written informed consent was obtained from all participants. Ethical approval was obtained from North West Centre for Research Ethics Committee (REC:99/8/84) and NRES Committee North East - County Durham and Tees Valley Ethics Committee (REC: 12/NE/0251). Experimental methods comply with the Helsinki Declaration.

Author details

[1]Arthritis Research UK Centre for Genetics and Genomics, Division of Musculoskeletal and Dermatological Sciences, School of Biological Sciences, Faculty of Biology, Medicine and Health, Manchester Academic Health Science Centre, The University of Manchester, Stopford Building, Oxford Road, Manchester M13 9PT, UK. [2]Nuclear Dynamics Programme, The Babraham Institute, Cambridge CB22 3AT, UK. [3]NIHR Manchester Musculoskeletal BRU, Manchester Academic Health Sciences Centre, Central Manchester Foundation Trust, Manchester, UK. [4]Institute of Cellular Medicine (Musculoskeletal Research Group), Newcastle University, Newcastle upon Tyne NE2 4HH, UK. [5]Center of Experimental Rheumatology Department of Rheumatology, University Hospital of Zurich, Wagistrasse 14, 8952 Schlieren, Switzerland.

Received: 26 May 2016 Accepted: 5 October 2016
Published online: 01 November 2016

References

1. Welter D, MacArthur J, Morales J, Burdett T, Hall P, Junkins H, et al. The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. Nucleic Acids Res. 2014;42:D1001–6.
2. Freedman ML, Monteiro AN, Gayther SA, Coetzee GA, Risch A, Plass C, et al. Principles for the post-GWAS functional characterization of cancer risk loci. Nat Genet. 2011;43:513–8.
3. Ward LD, Kellis M. Interpreting noncoding genetic variation in complex traits and human disease. Nat Biotechnol. 2012;30:1095–106.
4. Fairfax BP, Humburg P, Makino S, Naranbhai V, Wong D, Lau E, et al. Innate immune activity conditions the effect of regulatory variants upon monocyte gene expression. Science. 2014;343:1246949.
5. Farh KK, Marson A, Zhu J, Kleinewietfeld M, Housley WJ, Beik S, et al. Genetic and epigenetic fine mapping of causal autoimmune disease variants. Nature. 2015;518:337–43.
6. Ye CJ, Feng T, Kwon HK, Raj T, Wilson MT, Asinovski N, et al. Intersection of population variation and autoimmunity genetics in human T cell activation. Science. 2014;345:1254665.
7. Davison LJ, Wallace C, Cooper JD, Cope NF, Wilson NK, Smyth DJ, et al. Long-range DNA looping and gene expression analyses identify DEXI as an autoimmune disease candidate gene. Hum Mol Genet. 2012;21:322–33.
8. Pomerantz MM, Ahmadiyeh N, Jia L, Herman P, Verzi MP, Doddapaneni H, et al. The 8q24 cancer risk variant rs6983267 shows long-range interaction with MYC in colorectal cancer. Nat Genet. 2009;41:882–4.
9. Wang S, Wen F, Wiley GB, Kinter MT, Gaffney PM. An enhancer element harboring variants associated with systemic lupus erythematosus engages the TNFAIP3 promoter to influence A20 expression. PLoS Genet. 2013;9:e1003750.
10. Zhang X, Cowper-Sal lR, Bailey SD, Moore JH, Lupien M. Integrative functional genomics identifies an enhancer looping to the SOX9 gene disrupted by the 17q24.3 prostate cancer risk locus. Genome Res. 2012;22:1437–46.
11. Bulger M, Groudine M. Functional and mechanistic diversity of distal transcription enhancers. Cell. 2011;144:327–39.
12. Sanyal A, Lajoie BR, Jain G, Dekker J. The long-range interaction landscape of gene promoters. Nature. 2012;489:109–13.
13. Schoenfelder S, Furlan-Magaril M, Mifsud B, Tavares-Cadete F, Sugar R, Javierre BM, et al. The pluripotent regulatory circuitry connecting promoters to their long-range interacting elements. Genome Res. 2015;25:582–97.
14. Dekker J, Rippe K, Dekker M, Kleckner N. Capturing chromosome conformation. Science. 2002;295:1306–11.
15. Dostie J, Richmond TA, Arnaout RA, Selzer RR, Lee WL, Honan TA, et al. Chromosome Conformation Capture Carbon Copy (5C): a massively parallel solution for mapping interactions between genomic elements. Genome Res. 2006;16:1299–309.
16. Simonis M, Klous P, Splinter E, Moshkin Y, Willemsen R, de WE, et al. Nuclear organization of active and inactive chromatin domains uncovered by chromosome conformation capture-on-chip (4C). Nat Genet. 2006;38:1348–54.
17. Lieberman-Aiden E, van Berkum NL, Williams L, Imakaev M, Ragoczy T, Telling A, et al. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. Science. 2009;326:289–93.
18. Dryden NH, Broome LR, Dudbridge F, Johnson N, Orr N, Schoenfelder S, et al. Unbiased analysis of potential targets of breast cancer susceptibility loci by Capture Hi-C. Genome Res. 2014;24:1854–68.

19. Jager R, Migliorini G, Henrion M, Kandaswamy R, Speedy HE, Heindl A, et al. Capture Hi-C identifies the chromatin interactome of colorectal cancer risk loci. Nat Commun. 2015;6:6178.
20. Mifsud B, Tavares-Cadete F, Young AN, Sugar R, Schoenfelder S, Ferreira L, et al. Mapping long-range promoter contacts in human cells with high-resolution capture Hi-C. Nat Genet. 2015;47:598–606.
21. Martin P, McGovern A, Orozco G, Duffus K, Yarwood A, Schoenfelder S, et al. Capture Hi-C reveals novel candidate genes and complex long-range interactions with related autoimmune risk loci. Nat Commun. 2015;6:10069.
22. Bowes J, Budu-Aggrey A, Huffmeier U, Uebe S, Steel K, Hebert HL, et al. Dense genotyping of immune-related susceptibility loci reveals new insights into the genetics of psoriatic arthritis. Nat Commun. 2015;6:6046.
23. Coenen MJ, Trynka G, Heskamp S, Franke B, van Diemen CC, Smolonska J, et al. Common and different genetic background for rheumatoid arthritis and coeliac disease. Hum Mol Genet. 2009;18:4195–203.
24. Graham RR, Cotsapas C, Davies L, Hackett R, Lessard CJ, Leon JM, et al. Genetic variants near TNFAIP3 on 6q23 are associated with systemic lupus erythematosus. Nat Genet. 2008;40:1059–61.
25. Jostins L, Ripke S, Weersma RK, Duerr RH, McGovern DP, Hui KY, et al. Host-microbe interactions have shaped the genetic architecture of inflammatory bowel disease. Nature. 2012;491:119–24.
26. Nair RP, Duffin KC, Helms C, Ding J, Stuart PE, Goldgar D, et al. Genome-wide scan reveals association of psoriasis with IL-23 and NF-kappaB pathways. Nat Genet. 2009;41:199–204.
27. Onengut-Gumuscu S, Chen WM, Burren O, Cooper NJ, Quinlan AR, Mychaleckyj JC, et al. Fine mapping of type 1 diabetes susceptibility loci and evidence for colocalization of causal variants with lymphoid gene enhancers. Nat Genet. 2015;47:381–6.
28. Thomson W, Barton A, Ke X, Eyre S, Hinks A, Bowes J, et al. Rheumatoid arthritis association at 6q23. Nat Genet. 2007;39:1431–3.
29. Wellcome Trust Case Control Consortium. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature. 2007;447:661–78.
30. Adrianto I, Wen F, Templeton A, Wiley G, King JB, Lessard CJ, et al. Association of a functional variant downstream of TNFAIP3 with systemic lupus erythematosus. Nat Genet. 2011;43:253–8.
31. Musone SL, Taylor KE, Lu TT, Nititham J, Ferreira RC, Ortmann W, et al. Multiple polymorphisms in the TNFAIP3 region are independently associated with systemic lupus erythematosus. Nat Genet. 2008;40:1062–4.
32. Ward LD, Kellis M. HaploReg: a resource for exploring chromatin states, conservation, and regulatory motif alterations within sets of genetically linked variants. Nucleic Acids Res. 2012;40:D930–4.
33. Boyle AP, Hong EL, Hariharan M, Cheng Y, Schaub MA, Kasowski M, et al. Annotation of functional variation in personal genomes using RegulomeDB. Genome Res. 2012;22:1790–7.
34. Elsby LM, Orozco G, Denton J, Worthington J, Ray DW, Donn RP. Functional evaluation of TNFAIP3 (A20) in rheumatoid arthritis. Clin Exp Rheumatol. 2010;28:708–14.
35. Huber LC, Distler O, Tarner I, Gay RE, Gay S, Pap T. Synovial fibroblasts: key players in rheumatoid arthritis. Rheumatology (Oxford). 2006;45:669–75.
36. Hayden MS, West AP, Ghosh S. NF-kappaB and the immune response. Oncogene. 2006;25:6758–80.
37. Bours V, Franzoso G, Azarenko V, Park S, Kanno T, Brown K, et al. The oncoprotein Bcl-3 directly transactivates through kappa B motifs via association with DNA-binding p50B homodimers. Cell. 1993;72:729–39.
38. Carmody RJ, Ruan Q, Palmer S, Hilliard B, Chen YH. Negative regulation of toll-like receptor signaling by NF-kappaB p50 ubiquitination blockade. Science. 2007;317:675–8.
39. Wulczyn FG, Naumann M, Scheidereit C. Candidate proto-oncogene bcl-3 encodes a subunit-specific inhibitor of transcription factor NF-kappa B. Nature. 1992;358:597–9.
40. Schoenfelder S, Clay I, Fraser P. The transcriptional interactome: gene expression in 3D. Curr Opin Genet Dev. 2010;20:127–33.
41. Trynka G, Sandor C, Han B, Xu H, Stranger BE, Liu XS, et al. Chromatin marks identify critical cell types for fine mapping complex trait variants. Nat Genet. 2013;45:124–30.
42. Rao SS, Huntley MH, Durand NC, Stamenova EK, Bochkov ID, Robinson JT, et al. A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. Cell. 2014;159:1665–80.
43. Pestka S, Krause CD, Sarkar D, Walter MR, Shi Y, Fisher PB. Interleukin-10 and related cytokines and receptors. Annu Rev Immunol. 2004;22:929–79.
44. Rutz S, Wang X, Ouyang W. The IL-20 subfamily of cytokines–from host defence to tissue homeostasis. Nat Rev Immunol. 2014;14:783–95.
45. Su AI, Wiltshire T, Batalov S, Lapp H, Ching KA, Block D, et al. A gene atlas of the mouse and human protein-encoding transcriptomes. Proc Natl Acad Sci U S A. 2004;101:6062–7.
46. Corvaisier M, Delneste Y, Jeanvoine H, Preisser L, Blanchard S, Garo E, et al. IL-26 is overexpressed in rheumatoid arthritis and induces proinflammatory cytokine production and Th17 cell generation. PLoS Biol. 2012;10:e1001395.
47. Hsu YH, Li HH, Hsieh MY, Liu MF, Huang KY, Chin LS, et al. Function of interleukin-20 as a proinflammatory molecule in rheumatoid and experimental arthritis. Arthritis Rheum. 2006;54:2722–33.
48. Ikeuchi H, Kuroiwa T, Hiramatsu N, Kaneko Y, Hiromura K, Ueki K, et al. Expression of interleukin-22 in rheumatoid arthritis: potential role as a proinflammatory cytokine. Arthritis Rheum. 2005;52:1037–46.
49. Kragstrup TW, Otkjaer K, Holm C, Jorgensen A, Hokland M, Iversen L, et al. The expression of IL-20 and IL-24 and their shared receptors are increased in rheumatoid arthritis and spondyloarthropathy. Cytokine. 2008;41:16–23.
50. Sakurai N, Kuroiwa T, Ikeuchi H, Hiramatsu N, Maeshima A, Kaneko Y, et al. Expression of IL-19 and its receptors in RA: potential role for synovial hyperplasia formation. Rheumatology (Oxford). 2008;47:815–20.
51. Hsu YH, Hsieh PP, Chang MS. Interleukin-19 blockade attenuates collagen-induced arthritis in rats. Rheumatology (Oxford). 2012;51:434–42.
52. Ouyang W, Rutz S, Crellin NK, Valdez PA, Hymowitz SG. Regulation and functions of the IL-10 family of cytokines in inflammation and disease. Annu Rev Immunol. 2011;29:71–109.
53. Blumberg H, Conklin D, Xu WF, Grossmann A, Brender T, Carollo S, et al. Interleukin 20: discovery, receptor identification, and role in epidermal function. Cell. 2001;104:9–19.
54. He M, Liang P. IL-24 transgenic mice: in vivo evidence of overlapping functions for IL-20, IL-22, and IL-24 in the epidermis. J Immunol. 2010;184:1793–8.
55. Wolk K, Haugen HS, Xu W, Witte E, Waggie K, Anderson M, et al. IL-22 and IL-20 are key mediators of the epidermal alterations in psoriasis while IL-17 and IFN-gamma are not. J Mol Med (Berl). 2009;87:523–36.
56. Otkjaer K, Kragballe K, Funding AT, Clausen JT, Noerby PL, Steiniche T, et al. The dynamics of gene expression of interleukin-19 and interleukin-20 and their receptors in psoriasis. Br J Dermatol. 2005;153:911–8.
57. Romer J, Hasselager E, Norby PL, Steiniche T, Thorn CJ, Kragballe K. Epidermal overexpression of interleukin-19 and −20 mRNA in psoriatic skin disappears after short-term treatment with cyclosporine a or calcipotriol. J Invest Dermatol. 2003;121:1306–11.
58. Wolk K, Kunz S, Witte E, Friedrich M, Asadullah K, Sabat R. IL-22 increases the innate immunity of tissues. Immunity. 2004;21:241–54.
59. Gottlieb AB, Krueger JG, Sandberg LM, Gothberg M, Skolnick BE. First-in-human, phase 1, randomized, dose-escalation trial with recombinant anti-il-20 monoclonal antibody in patients with psoriasis. PLoS One. 2015;10:e0134703.
60. Senolt L, Leszczynski P, Dokoupilova E, Gothberg M, Valencia X, Hansen BB, et al. Efficacy and safety of anti-interleukin-20 monoclonal antibody in patients with rheumatoid arthritis: a randomized phase IIa trial. Arthritis Rheumatol. 2015;67:1438–48.
61. Andoh A, Shioya M, Nishida A, Bamba S, Tsujikawa T, Kim-Mitsuyama S, et al. Expression of IL-24, an activator of the JAK1/STAT3/SOCS3 cascade, is enhanced in inflammatory bowel disease. J Immunol. 2009;183:687–95.
62. Dambacher J, Beigel F, Zitzmann K, De Toni EN, Goke B, Diepolder HM, et al. The role of the novel Th17 cytokine IL-26 in intestinal inflammation. Gut. 2009;58:1207–17.
63. Fonseca-Camarillo G, Furuzawa-Carballeda J, Llorente L, Yamamoto-Furusho JK. IL-10– and IL-20–expressing epithelial and inflammatory cells are increased in patients with ulcerative colitis. J Clin Immunol. 2013;33:640–8.
64. Fonseca-Camarillo G, Furuzawa-Carballeda J, Granados J, Yamamoto-Furusho JK. Expression of interleukin (IL)-19 and IL-24 in inflammatory bowel disease patients: a cross-sectional study. Clin Exp Immunol. 2014;177:64–75.
65. Nelson MR, Tipney H, Painter JL, Shen J, Nicoletti P, Shen Y, et al. The support of human genetic evidence for approved drug indications. Nat Genet. 2015;47:856–60.
66. Hu X, Ivashkiv LB. Cross-regulation of signaling pathways by interferon-gamma: implications for immune responses and autoimmune diseases. Immunity. 2009;31:539–50.
67. Tang Q, Danila MI, Cui X, Parks L, Baker BJ, Reynolds RJ, et al. Expression of interferon-gamma receptor genes in peripheral blood mononuclear cells is associated with rheumatoid arthritis and its radiographic severity in African Americans. Arthritis Rheumatol. 2015;67:1165–70.

68. Barreiro LB, Tailleux L, Pai AA, Gicquel B, Marioni JC, Gilad Y. Deciphering the genetic architecture of variation in the immune response to Mycobacterium tuberculosis infection. Proc Natl Acad Sci U S A. 2012;109:1204–9.
69. Fairfax BP, Makino S, Radhakrishnan J, Plant K, Leslie S, Dilthey A, et al. Genetics of gene expression in primary immune cells identifies cell type-specific master regulators and roles of HLA alleles. Nat Genet. 2012;44:502–10.
70. Hu X, Kim H, Raj T, Brennan PJ, Trynka G, Teslovich N, et al. Regulation of gene expression in autoimmune disease loci and the genetic basis of proliferation in CD4+ effector memory T cells. PLoS Genet. 2014;10:e1004404.
71. Lee MN, Ye C, Villani AC, Raj T, Li W, Eisenhaure TM, et al. Common genetic variants modulate pathogen-sensing responses in human dendritic cells. Science. 2014;343:1246980.
72. Romanoski CE, Lee S, Kim MJ, Ingram-Drake L, Plaisier CL, Yordanova R, et al. Systems genetics analysis of gene-by-environment interactions in human cells. Am J Hum Genet. 2010;86:399–410.
73. Catrysse L, Vereecke L, Beyaert R. van LG. A20 in inflammation and autoimmunity. Trends Immunol. 2014;35:22–31.
74. Lee EG, Boone DL, Chai S, Libby SL, Chien M, Lodolce JP, et al. Failure to regulate TNF-induced NF-kappaB and cell death responses in A20-deficient mice. Science. 2000;289:2350–4.
75. Kilpinen H, Waszak SM, Gschwind AR, Raghav SK, Witwicki RM, Orioli A, et al. Coordinated effects of sequence variation on DNA binding, chromatin structure, and transcription. Science. 2013;342:744–7.
76. Dixon JR, Selvaraj S, Yue F, Kim A, Li Y, Shen Y, et al. Topological domains in mammalian genomes identified by analysis of chromatin interactions. Nature. 2012;485:376–80.
77. Pombo A, Dillon N. Three-dimensional genome architecture: players and mechanisms. Nat Rev Mol Cell Biol. 2015;16:245–57.
78. Kim YJ, Cecchini KR, Kim TH. Conserved, developmentally regulated mechanism couples chromosomal looping and heterochromatin barrier activity at the homeobox gene A locus. Proc Natl Acad Sci U S A. 2011;108:7391–6.
79. Duggal G, Wang H, Kingsford C. Higher-order chromatin domains link eQTLs with the expression of far-away genes. Nucleic Acids Res. 2014;42:87–96.
80. Lupianez DG, Kraft K, Heinrich V, Krawitz P, Brancati F, Klopocki E, et al. Disruptions of topological chromatin domains cause pathogenic rewiring of gene-enhancer interactions. Cell. 2015;161:1012–25.
81. Eyre S, Bowes J, Diogo D, Lee A, Barton A, Martin P, et al. High-density genetic mapping identifies new susceptibility loci for rheumatoid arthritis. Nat Genet. 2012;44:1336–40.
82. Hinks A, Cobb J, Marion MC, Prahalad S, Sudman M, Bowes J, et al. Dense genotyping of immune-related disease regions identifies 14 new susceptibility loci for juvenile idiopathic arthritis. Nat Genet. 2013;45:664–9.
83. Okada Y, Wu D, Trynka G, Raj T, Terao C, Ikari K, et al. Genetics of rheumatoid arthritis contributes to biology and drug discovery. Nature. 2014;506:376–81.
84. Hughes JR, Roberts N, McGowan S, Hay D, Giannoulatou E, Lynch M, et al. Analysis of hundreds of cis-regulatory landscapes at high resolution in a single, high-throughput experiment. Nat Genet. 2014;46:205–12.
85. Belton JM, McCord RP, Gibcus JH, Naumova N, Zhan Y, Dekker J. Hi-C: a comprehensive technique to capture the conformation of genomes. Methods. 2012;58:268–76.
86. van Berkum NL, Lieberman-Aiden E, Williams L, Imakaev M, Gnirke A, Mirny LA, et al. Hi-C: a method to study the three-dimensional architecture of genomes. J Vis Exp. 2010. DOI: 10.3791/1869.
87. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics. 2014;30:2114–20.
88. Naumova N, Smith EM, Zhan Y, Dekker J. Analysis of long-range chromatin interactions using Chromosome Conformation Capture. Methods. 2012;58:192–203.
89. Zhou X, Maricque B, Xie M, Li D, Sundaram V, Martin EA, et al. The Human Epigenome Browser at Washington University. Nat Methods. 2011;8:989–90.
90. Zhou X, Lowdon RF, Li D, Lawson HA, Madden PA, Costello JF, et al. Exploring long-range genome interactions using the WashU Epigenome Browser. Nat Methods. 2013;10:375–6.
91. Montgomery SB, Sammeth M, Gutierrez-Arcelus M, Lach RP, Ingle C, Nisbett J, et al. Transcriptome genetics using second generation sequencing in a Caucasian population. Nature. 2010;464:773–7.
92. Stranger BE, Nica AC, Forrest MS, Dimas A, Bird CP, Beazley C, et al. Population genomics of human gene expression. Nat Genet. 2007;39:1217–24.
93. Schadt EE, Molony C, Chudin E, Hao K, Yang X, Lum PY, et al. Mapping the genetic architecture of gene expression in human liver. PLoS Biol. 2008;6:e107.
94. Gibbs JR, van der Brug MP, Hernandez DG, Traynor BJ, Nalls MA, Lai SL, et al. Abundant quantitative trait loci exist for DNA methylation and gene expression in human brain. PLoS Genet. 2010;6:e1000952.
95. Li Q, Stram A, Chen C, Kar S, Gayther S, Pharoah P, et al. Expression QTL-based analyses reveal candidate causal genes and loci across five tumor types. Hum Mol Genet. 2014;23:5294–302.
96. Westra HJ, Peters MJ, Esko T, Yaghootkar H, Schurmann C, Kettunen J, et al. Systematic identification of trans eQTLs as putative drivers of known disease associations. Nat Genet. 2013;45:1238–43.
97. Hao K, Bosse Y, Nickle DC, Pare PD, Postma DS, Laviolette M, et al. Lung eQTLs to help reveal the molecular underpinnings of asthma. PLoS Genet. 2012;8:e1003029.
98. Koopmann TT, Adriaens ME, Moerland PD, Marsman RF, Westerveld ML, Lal S, et al. Genome-wide identification of expression quantitative trait loci (eQTLs) in human heart. PLoS One. 2014;9:e97380.
99. Grundberg E, Adoue V, Kwan T, Ge B, Duan QL, Lam KC, et al. Global analysis of the impact of environmental perturbation on cis-regulation of gene expression. PLoS Genet. 2011;7:e1001279.
100. Lappalainen T, Sammeth M, Friedlander MR, 't Hoen PA, Monlong J, Rivas MA, et al. Transcriptome and genome sequencing uncovers functional variation in humans. Nature. 2013;501:506–11.
101. Ramasamy A, Trabzuni D, Guelfi S, Varghese V, Smith C, Walker R, et al. Genetic variability in the regulation of gene expression in ten regions of the human brain. Nat Neurosci. 2014;17:1418–28.
102. Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res. 2015;43:e47.

Publication 6: Identifying Causal Genes at the Multiple Sclerosis Associated Region 6q23 Using Capture Hi-C

RESEARCH ARTICLE

Identifying Causal Genes at the Multiple Sclerosis Associated Region 6q23 Using Capture Hi-C

Paul Martin[1☯], Amanda McGovern[1☯], Jonathan Massey[1], Stefan Schoenfelder[2], Kate Duffus[1], Annie Yarwood[1], Anne Barton[1,3], Jane Worthington[1,3], Peter Fraser[2], Stephen Eyre[1], Gisela Orozco[1]*

1 Arthritis Research UK Centre for Genetics and Genomics, Centre for Musculoskeletal Research, Division of Musculoskeletal and Dermatological Siences, Faculty of Biology, Medicine and Health, Manchester Academic Health Science Centre, The University of Manchester, Stopford Building, Oxford Road, Manchester, M13 9PT, United Kingdom, **2** Nuclear Dynamics Programme, The Babraham Institute, Cambridge CB22 3AT, United Kingdom, **3** NIHR Manchester Musculoskeletal BRU, Manchester Academic Health Sciences Centre, Central Manchester Foundation Trust, Manchester, United Kingdom

☯ These authors contributed equally to this work.
* gisela.orozco@manchester.ac.uk

OPEN ACCESS

Citation: Martin P, McGovern A, Massey J, Schoenfelder S, Duffus K, Yarwood A, et al. (2016) Identifying Causal Genes at the Multiple Sclerosis Associated Region 6q23 Using Capture Hi-C. PLoS ONE 11(11): e0166923. doi:10.1371/journal.pone.0166923

Editor: David R. Booth, Westmead Millennium Institute for Medical Research, AUSTRALIA

Received: July 7, 2016

Accepted: November 6, 2016

Published: November 18, 2016

Data Availability Statement: Raw data and HindIII restriction fragment interaction counts are available in the NCBI Gene Expression Omnibus (GEO; http://www.ncbi.nlm.nih.gov/geo/) under accession number GSE69600.

Funding: Wellcome Trust Research Career Development Fellowship (095684); Arthritis Research UK (grant numbers 20385, 20571); Wellcome Trust (097820/Z/11/B); European Union's FP7 Health Programme (FP7-HEALTH-F2-2012-305549, Euro-TEAM, FP7/2007-2013); Innovative Medicines Initiative (BeTheCure project

Abstract

Background

The chromosomal region 6q23 has been found to be associated with multiple sclerosis (MS) predisposition through genome wide association studies (GWAS). There are four independent single nucleotide polymorphisms (SNPs) associated with MS in this region, which spans around 2.5 Mb. Most GWAS variants associated with complex traits, including these four MS associated SNPs, are non-coding and their function is currently unknown. However, GWAS variants have been found to be enriched in enhancers and there is evidence that they may be involved in transcriptional regulation of their distant target genes through long range chromatin looping.

Aim

The aim of this work is to identify causal disease genes in the 6q23 locus by studying long range chromatin interactions, using the recently developed Capture Hi-C method in human T and B-cell lines. Interactions involving four independent associations unique to MS, tagged by rs11154801, rs17066096, rs7769192 and rs67297943 were analysed using Capture Hi-C Analysis of Genomic Organisation (CHiCAGO).

Results

We found that the pattern of chromatin looping interactions in the MS 6q23 associated region is complex. Interactions cluster in two regions, the first involving the rs11154801 region and a second containing the rs17066096, rs7769192 and rs67297943 SNPs. Firstly, SNPs located within the *AHI1* gene, tagged by rs11154801, are correlated with expression of *AHI1 and* interact with its promoter. These SNPs also interact with other potential

115142); Medical Research Council (MR/K015346/1); National Institute for Health Research Manchester Musculoskeletal Biomedical Research Unit; Biotechnology and Biological Sciences Research Council UK (BBS/E/B/000C0405).

Competing Interests: The authors have declared that no competing interests exist.

candidate genes such as *SGK1* and *BCLAF1*. Secondly, the rs17066096, rs7769192 and rs67297943 SNPs interact with each other and with immune-related genes such as *IL20RA*, *IL22RA2*, *IFNGR1* and *TNFAIP3*. Finally, the above-mentioned regions interact with each other and therefore, may co-regulate these target genes.

Conclusion

These results suggest that the four 6q23 variants, independently associated with MS, are involved in the regulation of several genes, including immune genes. These findings could help understand mechanisms of disease and suggest potential novel therapeutic targets.

Introduction

Genome wide association studies (GWAS) have been pivotal in identifying genetic associations with single nucleotide polymorphisms (SNPs) in many complex diseases, including multiple sclerosis (MS) [1–4]. MS is an inflammatory demyelinating disease of the central nervous system (CNS) and is a common cause of chronic neurological disability, showing moderate heritability (λ_s ~6.3) [5]. Similar to many autoimmune diseases, the major histocompatibility complex (MHC) represents the largest single genetic risk factor for MS, with multiple non-HLA loci, discovered in large international GWAS, contributing smaller individual effects to disease susceptibility. Due to the extensive overlap of genetic loci between multiple autoimmune diseases, the International Multiple Sclerosis Genetics Consortium (IMSGC) conducted a large study using the Illumina Immunochip genotyping array, identifying 48 new and validating 49 previously discovered non-MHC susceptibility variants for MS [6]. Among these variants, four mapped to the 6q23 region, which is also associated with other autoimmune diseases including rheumatoid arthritis (RA), systemic lupus erythematosus (SLE), celiac disease (CeD), type 1 diabetes (T1D), inflammatory bowel disease (IBD), psoriasis (Ps) and psoriatic arthritis (PsA), and containing several candidate genes, such as *TNFAIP3*, *AHI1* and *IL22RA2* [7–13].

The 6q23 locus, like many other GWAS loci, shows extensive overlap with many other autoimmune diseases and demonstrates a complex pattern of different associations attributable to different diseases. This sharing of associated loci led to the Immunochip array which contains three regions densely mapped and capturing four independent associations with MS (**Table 1 and Fig 1**) in the 6q23 region. The first, tagged by rs11154801, is located within an intron of the *AHI1* gene required for both cerebellar and cortical development. The second region, tagged by rs17066096, is an intergenic region 87kb 5' of *IL20RA* and 12kb 3' of *IL22RA2*. The third region covers 430kb, encompassing a *PTPN11* pseudogene (*RP11-95M15.2*), *TNFAIP3* and several lncRNAs and contains two independent associations (rs7769192 & rs67297943). Interestingly, while other SNP associations are shared between autoimmune diseases, the MS associated SNPs are unique to MS alone (**S1 Data**). As such, these MS associated SNPs could offer an insight into the mechanisms affecting MS at this locus.

However, due to the design of GWAS, these lead genetic associations do not necessarily represent the causal variant but instead a number of variants in strong linkage disequilibrium with them. In addition, associated SNPs have generally been annotated to the closest, most biologically plausible gene. Evidence suggests that GWAS discovered SNPs in general, including these associations within 6q23, are enriched in cell-type specific enhancer regions [14,15] which can regulate gene expression. Additionally, an individual's genotype can influence this

Table 1. MS 6q23 Immunochip associated regions.

Region Co-ordinates (GRCh37)			Index SNP	Reported Gene	MAF	P	OR
Chr.	**Start**	**End**					
6	134946991	135498875	rs11154801	*AHI1*	0.37	1.80 x 10^{-20}	1.12
6	136685539	136875216	rs17066096	*IL22RA2*	0.23	1.60 x 10^{-23}	1.14
6	137231416	137660037	rs7769192[a]	-	0.45	3.30 x 10^{-09}	1.08
			rs67297943	*TNFAIP3*	0.22	5.50 x 10^{-13}	1.11

All P values are from the joint analysis by Beecham et al.[6]. Chr., chromosome; MAF, minor allele frequency; OR, odds ratio.

a, P values and ORs shown after conditioning on rs67297943.

doi:10.1371/journal.pone.0166923.t001

expression (expression quantitative trait loci (eQTL)), potentially leading to disease. It has been shown that enhancers can regulate genes located some distance away through long-range chromatin interactions [16]. Therefore, confidently assigning causal SNPs, genes and cell types to these and other GWAS signals remains a major challenge. Potential long-range interactions have previously been prohibitive to investigate as methods, such as 3C and Hi-C, required interacting regions to be considered *a priori* or, lacked throughput and resolution. Capture Hi-C was developed to overcome these limitations by enriching a Hi-C library using RNA baits designed to specific restriction fragments. This approach reduces library complexity, increases power and subsequently allows the identification of statistically significant chromatin interactions at a restriction fragment resolution (~4kb). As part of a large study investigating the interactions with associated regions in four autoimmune diseases [17], several sites within the 6q23 region were targeted, including associated regions and promoters of nearby genes (**Table 1** and **Fig 1**). Our Capture Hi-C data represents a unique opportunity to explore this region for MS and offer an insight into the mechanisms specifically affecting MS at this locus and how they compare with other autoimmune diseases. The aim of this study was to use this chromatin interaction experiment to explore the unique genetic associations with MS in the 6q23 region to identify possible target causal genes whose expression could be perturbed in at risk individuals. The ultimate goal is to help translate GWAS findings into clinical benefit, as the identification of causal genes can pinpoint biological mechanisms altered in disease and suggest potential therapeutic targets or drug repositioning.

We demonstrate that the MS associated region 6q23 presents numerous, complex chromatin looping interactions clustered in two regions. The first contains SNPs located within the *AHI1* gene, tagged by rs11154801, and correlated with expression, which interact with the

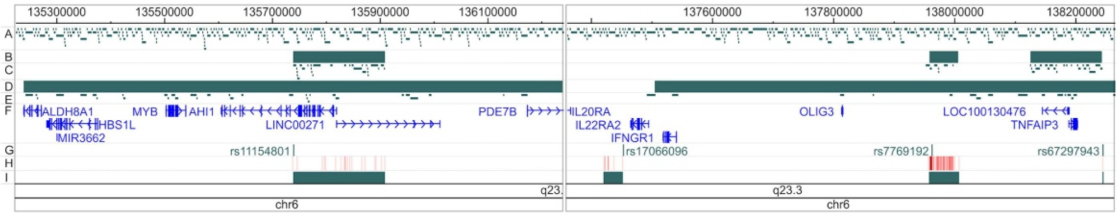


Fig 1. Overview of MS 6q23 Immunochip associated regions. Tracks are labelled as follows: A–HindIII restriction fragments; B–LD regions targeted in 'region' Capture Hi-C; C–HindIII restriction fragments targeted in 'region' Capture Hi-C; D–Gene regions targeted in 'promoter' Capture Hi-C; E–HindIII restriction fragments targeted in 'promoter' Capture Hi-C; F–RefSeq genes (packed for clarity); G–MS index SNPs; H–Density of MS LD SNPs ($r^2 \geq 0.8$) and I–MS LD regions. The genomic region chr6:136,238,000–137,360,000 has been omitted for clarity. All co-ordinates are based on GRCh37. Generated using the WashU EpiGenome Browser (http://epigenomegateway.wustl.edu/browser/).

doi:10.1371/journal.pone.0166923.g001

AHI1 promoter thereby supporting the gene candidature of *AHI1*. Interestingly, these SNPs also interact with other potential candidate genes such as *SGK1* and *BCLAF1*, suggesting they may regulate multiple loci. The second region encompasses the rs17066096, rs7769192 and rs67297943 associated regions and interact with each other and with immune-related genes, such as *IL20RA*, *IL22RA2*, *IFNGR1* and *TNFAIP3*. Additionally, these regions interact with each other and therefore, may co-regulate these target genes.

Materials and Methods

MS SNP Associations & Regions

All MS SNP associations in the 6q23 region were taken from the IMSGC Immunochip study [6]. All SNPs in linkage disequilibrium (LD) ($r2 \geq 0.8$) with each lead Immunochip MS SNP were identified using European samples from the 1000 Genomes Phase 3 release. Associated regions for each lead association were defined by the two terminal SNPs in LD.

Cell culture

B-lymphoblastoid cell lines (LCL) were obtained directly from Coriell Institute for Medical Research (catalogue number GM12878). Cells were grown in vented 25cm^2 cell culture flasks containing 10-20mls of Roswell Park Memorial Institute (RPMI)-1640 + 2mM L-glutamine culture medium, supplemented with 15% foetal bovine serum (FBS). Flasks were incubated upright at 37°C/5% CO^2. Cultures were regularly monitored to maintain a cell density between $2{\times}10^5$–$5{\times}10^5$ viable cells/ml. Cells were split every 2 days into fresh medium until they reached a maximum density of $1{\times}10^6$ cells/ml.

Jurkat E6.1 human leukaemic T-lymphoblast cells were obtained directly from LGC Standards (catalogue number ATCC® TIB-152™). Cells were grown in vented 25cm^2 cell culture flasks containing 10-20mls of RPMI-1640 + 2mM L-glutamine, supplemented with 10% FBS. Flasks were incubated upright at 37°C/5% CO$_2$ and the cultures regularly monitored to maintain a cell density between $3{\times}10^5$–$9{\times}10^5$ viable cells/ml.

These cell lines are not listed in the in the database of commonly misidentified cell lines maintained by ICLAC, were authenticated using STR analysis and were tested for mycoplasma contamination (MycoSEQ® Mycoplasma Detection System, 4460625, Life Technologies).

Capture Hi-C

Capture Hi-C data was produced as part of a larger study targeting all regions associated with four autoimmune diseases (RA, JIA, PsA and T1D) and separately, all promoters within these regions [17]. Briefly, all promoters within 1Mb of associated SNPs were selected and RNA baits were designed to the ends of all fragments within 500bp of the transcription start sites. Separately, associated regions were defined by SNPs in LD ($r^2 \geq 0.8$) and all restriction fragments not selected for the promoter capture experiment were targeted. Experiments were performed using human T-cell (Jurkat) and B-cell (GM12878) lines. Capture Hi-C libraries were sequenced using 75bp paired-end reads on an Illumina HiSeq 2500. Resulting reads were mapped to restriction fragments and filtered using the Hi-C User Pipeline (HICUP http://www.bioinformatics.babraham.ac.uk/projects/hicup). Chromatin interactions were analysed using CHiCAGO (Capture Hi-C Analysis Of Genomic Organisation [18], http://regulatorygenomicsgroup.org/chicago), a publicly available, open-source, bespoke statistical model for detecting significant interactions in Capture Hi-C data at a single restriction fragment resolution. Further filtering was carried out using the BEDTools v2.21.0 pairtobed command to identify significant interactions involving the MS associated regions.

Chromatin interactions identified in the Capture Hi-C data were further validated against dense Hi-C data generated by Rao *et al.* [19] in GM12878 cells. No data was available for the Jurkat T-cell line. Raw contact matrices and normalisation matrices for GM12878 cells at 5kb resolution were obtained from GEO accession GSE63525. Observed and expected contact matrices were normalised using the Knight and Ruiz normalisation matrices as described in the accompanying documentation. Observed/expected (O/E) values were calculated and further filtered by O/E $\geq$5 and normalised read count $\geq$ 5. BEDTOOLS was used to obtain the overlap of interactions observed in our data and the Rao *et al.* [19] data.

Expression quantitative trait loci (eQTLs) analysis

Publicly available datasets from Westra, *et al.* [20], the GEUVADIS analysis (http://www.ebi.ac.uk/arrayexpress/files/E-GEUV-1/analysis_results/) and Raj *et al.* [21] were queried directly. Additional datasets were also queried through HaploReg [22]. Two whole-genome gene expression datasets were also available in-house: CD4+ and CD8+ T-cells from 21 healthy individuals of the National Repository of Healthy Volunteers (NRHV), The University of Manchester. Written informed consent was obtained from all subjects. Ethical approval was obtained from North West Centre for Research Ethics Committee (REC: 99/8/84). Samples were all of Caucasian ancestry with a median age of 50.5 years (26–82 years) and comprised of 8 males and 13 females. mRNA was isolated from sorted cell subsets, quality and concentration assessed using the Agilent Bioanalyzer and Nanodrop, before cDNA/cRNA conversion using Illumina TotalPrep RNA Amplification Kits. 750ng of cRNA was hybridised to HumanHT-12 v4 Expression BeadChip arrays according to the manufacturer's protocol before being scanned on the Illumina iScan system. Raw expression data were exported from Illumina GenomeStudio and analysed using the R Bioconductor package 'limma' 81. Briefly, the neqc function was used for log2 transformation of the data, background correction and quantile normalisation using control probes. Principal Component Analysis was used to detect batch effects. The cDNA/cRNA conversion produced the largest batch effect in both cohorts and was corrected using ComBat (in R Bioconductor package sva) (http://bioconductor.org/packages/release/bioc/html/sva.html). Genome-wide genotype data was generated using the Illumina HumanCoreExome BeadChip kit. Genotype data was aligned to the 1000 genomes reference strand, pre-phased using SHAPEIT2 (v2.r727), before imputation using IMPUTE2 (v2.3.0) with the 1000 genome reference panel Phase 1. Imputed data was hard-called to genotypes using an INFO score cut-off of 0.8 and posterior probability of 0.9. The effect of the SNPs on gene expression was analysed using MatrixEQTL (v 2.1.0) (http://www.bios.unc.edu/research/genomic_software/Matrix_eQTL/) with an additive linear model. Only SNPs within 4Mb of a gene expression probe were considered to be cis-eQTL.

Bioinformatics Refinement of SNPs

SNPs were annotated using data from HaploReg v4.1 [22] and RegulomeDB v1.1 [23] for each LD SNP and combined with our Capture Hi-C data. SNPs attaining a RegulomeDB score of $\geq$5 and showing evidence of chromatin interactions in either cell type were selected as potentially causal and merit further investigation.

Results and Discussion

Capture Hi-C

Chromatin interactions at the 6q23 locus were analysed as part of a larger study that included all known risk loci for RA, juvenile idiopathic arthritis (JIA), PsA and T1D. We performed two

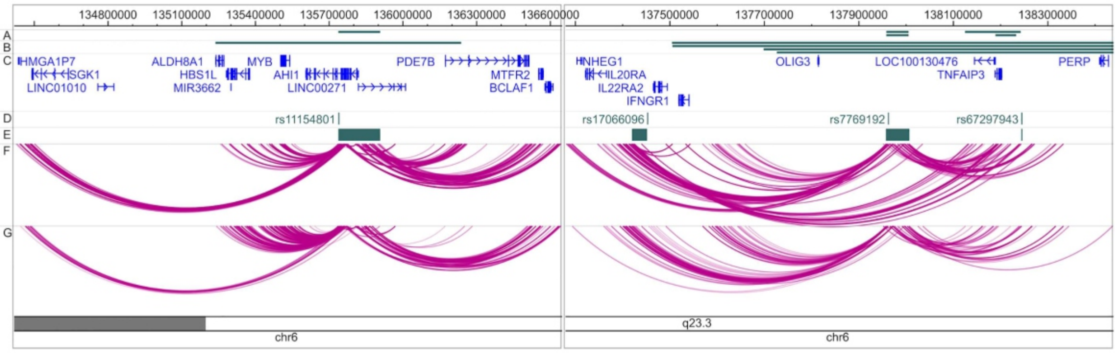


Fig 2. Overview of MS 6q23 interactions. Tracks are labelled as follows: A–LD regions targeted in 'region' Capture Hi-C; B–Gene regions targeted in 'promoter' Capture Hi-C; C–RefSeq genes (packed for clarity); D–MS index SNPs; E–MS LD regions; F–Interactions observed in the GM12878 B-cell line and G–Interactions observed in the Jurkat T-cell line. Promoter and region Capture Hi-C experiments have been merged for clarity. The genomic region chr6:136,650,000–137,280,000 has been omitted for clarity. All co-ordinates are based on GRCh37. Generated using the WashU EpiGenome Browser (http://epigenomegateway.wustl.edu/browser/).

doi:10.1371/journal.pone.0166923.g002

different Capture Hi-C experiments: firstly, the Region Capture targeted the LD regions (r^2>0.8) for all SNPs associated with each disease (Fig 1B); secondly, the Promoter Capture targeted all known gene promoters overlapping a region 500kb upstream and downstream of the lead disease associated SNP (Fig 1D). Capture Hi-C libraries were generated for two cell lines: GM12878, a B-lymphoblastoid cell line, and Jurkat, a CD4+ T-lymphoblastoid cell line.

Our Capture Hi-C experiments revealed that the 6q23 region presents a complex pattern of chromosomal interactions, highlighting both new and previously implicated genes for disease risk. Overall, 827 unique interactions involving MS 6q23 associated regions were observed across both cell lines and both capture experiments (promoter & region) (Fig 2 and S1 Fig). Each cell line demonstrated similar interaction patterns in both capture experiments. Encouragingly, there was a high degree of support from previously published Hi-C data obtained in GM12878 cells [19], with between 81% and 90% of interactions identified through our Capture Hi-C also being seen at an observed/expected ratio of ≥5 in previously published data.

The numerous chromosomal interactions detected appeared to cluster in two genomic locations, involving the rs11154801 region and a region containing the other three independent MS associations: rs17066096, rs7769192 and rs67297943 (Fig 2).

The rs11154801 LD block spans 170.4kb and contains several enhancers overlapping SNPs in strong LD with the lead association. It demonstrates a complex array of several long-range (>100kb) as well as shorter (<100kb) chromatin interactions (Fig 2 and S2 Fig) in both cell lines. Shorter internal chromatin interactions include ones with restriction fragments flanking the *AHI1* gene (previously assigned as the candidate gene to this variant) promoter, and SNPs in LD with rs11154801. We show how these SNPs, within the introns of *AHI1*, interact with the promoter region thereby supporting the *AHI1* hypothesis in both cell lines. Mutations in *AHI1* have been shown to cause Joubert syndrome [24], an autosomal recessive neurological condition causing symptoms including neonatal breathing abnormalities and mental retardation. Furthermore, it has been suggested that *AHI1* is required for both cerebellar and cortical development in humans and is expressed in the brain [25].

However, this locus may be more complex than previously thought, as long-range chromatin interactions, although more numerous in B-cells, were observed between the enhancer region and other compelling candidates such as *SGK1* and *BCLAF1* in both cell lines. The interaction with *SGK1* represents a >1.2Mb interaction and *Sgk1* knockout mice have been shown to have a reduced incidence of disease severity in experimental autoimmune encephalomyelitis (EAE), a mouse model of MS [26]. Other long-range interactions include those to the *MYC* and *PDE7B* gene regions.

Finally, rs11154801 also showed an interaction in both B and T-cells with a region encompassing the promoters of *BCLAF1*, *MTFR2*, an antisense gene (*RP13-143G15.4*) overlapping *PDE7B* and a lincRNA (RP3-406A7.7). The *BCLAF1* gene encodes a transcriptional repressor which interacts with BCL2-family proteins, is expressed in the brain and overexpression can lead to cell apoptosis [27]. In lymphocytes, dysregulation of BCL2-family proteins has been shown to lead to a reduction of pro-apoptotic BCL-2 members and survival of T-cells in MS [28]. Additionally, *BCLAF1* has also been shown to be crucial in the homeostasis of T- and B-cell lineages and proliferation of T-cells [29]. These interactions suggest that the associations at this locus may have different or additional effects on disease than just the previously assigned *AHI1* gene.

The second cluster of chromatin interactions involve the remaining MS associated SNPs in the 6q23 region, rs17066096, rs7769192 and rs67297943. Our Capture Hi-C results showed that these three SNPs interact with each other and also with several genes with immune function such as *IL20RA*, *IL22RA2*, *IFNGR1* and *TNFAIP3*, suggesting that these variants may be involved in the regulation of common immune pathways (**Fig 2**).

The first two interacting regions in this cluster, identified by the promoter Capture Hi-C experiment and tagged by rs17066096, were only observed in B-cells and involved the *IL22RA2* and *IFNGR1* gene promoters, almost 71kb and 113kb away, respectively. The restriction fragment overlapping the rs17066096 LD block shows evidence of enhancer activity, as predicted by ChromHMM, and the furthest 5' SNP in LD with the index SNP is located within this enhancer.

The region tagged by rs7769192 spans 50kb and contains 72 SNPs in LD, 21 of which are perfectly correlated with the lead SNP. It shows evidence of multiple enhancers and in addition to the previously mentioned interaction with the rs17066096 region, interacts with five other regions, in both cell lines. The first of these is located >500kb 5' of rs7769192 and contains the *IL20RA* gene. The next two regions, located over 400kb away, are shared with the rs17066096 region and contain the *IL22RA2* and *IFNGR1* genes, providing further evidence of the interplay between the two associated regions. The product of the *IFNGR1* gene is a subunit of the interferon gamma (IFN-γ) receptor whose ligand, IFN-γ, is important in adaptive immunity and has been linked to many different autoimmune diseases [30]. The *IL20RA* and *IL22RA2* genes both encode receptors for members of the IL-20 sub-family of cytokines and both exhibit a pro-inflammatory effect [31]. Additionally, anti-IL20 therapy has recently been shown to be effective in the treatment of RA and psoriasis [32,33]. Although anti-IL20 therapy was not developed as a result of Capture Hi-C, the discovery of interactions between these genes and autoimmune disease associations, demonstrates the power of this technique to inform drug discovery or repositioning. Further evidence for anti-IL20 therapy comes from the interaction with *IL22RA2*. *IL22RA2* encodes a soluble receptor which binds to and inhibits IL-22, a cytokine which can stimulate pro-inflammatory epithelial defence mechanisms [34], preventing the interaction with its cell surface receptor. This evidence suggests that blocking the IL-20 pathway may be effective in the treatment of MS and other autoimmune diseases.

The fourth region is located 178kb 3' of the associated region and contains multiple promoters of the *TNFAIP3* gene as well as a non-coding processed transcript (RP11-356I2.4) of unknown function. The role of *TNFAIP3* in autoimmunity is well established and the gene product A20 is a protein that is induced by tumour necrosis factor (TNF) and inhibits NFκB activation and TNF-mediated apoptosis [35]. This locus within the 6q23 region is one of the most important autoimmunity risk loci, as it contains multiple SNPs strongly associated with many autoimmune diseases, including MS, RA, SLE, CeD, IBD, psoriasis and PsA, among others. Variants associated with most autoimmune diseases map to the *TNFAIP3* gene or its vicinity, including the MS SNPs rs17066096, rs7769192 and rs67297943.

The rs67297943 SNP is located within a predicted enhancer element in B-cells and also in a 48.8kb region showing multiple enhancer marks in both B and T-cells. Furthermore, the adjacent restriction fragment to rs67297943 interacts with the *IL20RA* promoter region only in B-cells in this experiment.

Our Capture Hi-C results suggest that rs17066096, rs7769192 and rs67297943 physically interact with several immune genes with pro-inflammatory roles, such as *IL20RA*, *IL22RA2*, *IFNGR1* and *TNFAIP3*, indicating that they may be involved in the inflammatory processes that typically occurs in autoimmunity. Conversely, rs11154801, which is exclusively associated to MS and not other autoimmune diseases, interacts with genes with neurological function, like *AHI1*, *SGK1* and *BCLAF1*. Intriguingly, these two separate regions, over 2.3Mb apart, interact with each other in T-cells but not B-cells (**S1 Fig**), suggesting that these pathways may converge to give rise to disease-specific MS mechanisms in a stimulus and cell type specific manner. In this regard, it has been previously shown that there is a correlation between chromatin interactions and gene co-expression [36–39] and it has been hypothesised that multiple co-regulated genes can interact and share regulatory elements at specialised 'transcription factories' [16]. Our data possibly supports this idea and suggests a possible co-regulation of genes in this region in MS.

It could be argued that the differences observed between cell types for the rs1154801 region and the rs17066096 region could also be attributable to genotype differences in the cell lines (**Table 2**). While the overall pattern of chromatin interactions is similar for the rs11154801, rs17066096 and MYB regions, the intensity of interactions observed does vary between cell lines and could be due to carriage of risk alleles in the B-cell line, absent in the T-cell line. It is therefore important to validate the chromatin interactions in a genotype specific manner.

Expression Quantitative Trait Loci (eQTLs)

Public databases were interrogated for evidence of eQTLs for MS associated SNPs in the 6q23 region (rs11154801, rs17066096, rs7769192 and rs67297943) and all SNPs in LD ($r2>0.8$) with them. The SNPs within the intergenic region 5' of the *AHI1* gene, tagged by rs11154801 and interacting with the *AHI1* gene promoter, are correlated with *AHI1* mRNA expression in multiple tissues, including brain, nerve and whole blood. This further supports that *AHI1* is one of the causal genes within the 6q23 region.

Although many of the interactions detected in the Capture Hi-C experiment are between regions which show enhancer activity and regions which show active transcription, no eQTL evidence from public databases was observed for any of the other MS associated SNPs

Table 2. Associated SNP genotypes for GM12878 (B) and Jurkat (T) cell lines.

SNP	Risk Allele	GM12878		Jurkat	
		Genotype	Number of risk alleles	Genotype	Number of risk alleles
rs11154801	A	CA	1	CC	0
rs17066096	G	AA	0	AG	1
rs7769192	A	AA	2	unknown*	
rs67297943	C	TC	1	unknown*	

*Neither directly genotyped nor suitable proxies available on array

doi:10.1371/journal.pone.0166923.t002

investigated, other than rs11154801. This could be due to the distance cut-offs used to define *cis* effects or the selection of the correct cell type and stimulatory conditions, as eQTLs are known to be highly cell type and stimulus specific.

However, using in-house eQTL data on CD4+ and CD8+ primary T-cells from healthy donors, rs17066096 was shown to be correlated with expression of *IL20RA* in CD8+ T-cells ($P = 0.01$, **Fig 3**). Although the design of the Capture Hi-C experiment did not allow for the testing of chromatin interactions between the rs17066096 region and *IL20RA*, this eQTL suggests *IL20RA* could be important in the pathogenesis of MS in CD8+ T-cells, which have previously been shown to contribute to disease [40]. This data also confirmed the eQTL between rs11154801 and *AHI1* in both CD4+ and CD8+ T-cells (P = $2.0\text{x}10^{-4}$ and 0.02 respectively). Full eQTL results for all MS SNPs are presented in **S2 Data**. No other eQTLs were identified between MS SNPs and genes showing chromatin interactions, this may again be due to the selection of the correct cell type and stimulatory conditions.

Bioinformatics Refinement of SNPs

By utilising publically available data on regulatory elements obtained through HaploReg and RegulomeDB and augmenting with our Capture Hi-C data, we were able to refine large numbers of potential causal SNPs for three out of the four MS associated regions, by strong evidence of being in both a relevant cell-type enhancer region and interacting with a gene promoter (**Table 3** and **S3 Data**). For the rs11154801 region, 6 SNPs were identified out of 19 potential candidates; for the rs17066096, 3 SNPs were selected from a total of 7 in LD with the index association; and finally, for the rs7769192 region, we refined the potential candidates to 4 SNPs out of 72 in LD with rs7769192. No SNPs were found in LD with rs67297943 and although this SNP shows enhancer marks in 6 tissues, no interactions were identified with this region and as such further refinement was not possible. It will be imperative to follow-up putative SNPs and genes with functional assays and demonstrate their contribution to disease in relevant cell types in a biological context using genome editing techniques.

Conclusion

In conclusion, our work has strengthened the case for the *AHI1* gene candidate but also identified other potential MS gene targets, such as *SGK1*, *BCLAF1 IL20RA*, *IL22RA2*, *IFNGR1* and *TNFAIP3*. Additionally, we have shown a possible co-regulation of MS GWAS associations in the 6q23 region, which could help elucidate the pathogenesis of MS as well as other autoimmune diseases. These targets require further functional investigation which has been informed by the bioinformatics analysis. While the MS associations show evidence of interacting with other genes with no obvious role in MS pathogenesis, it is likely that they share regulatory elements within this region. It is however important to investigate these interactions, in addition to ones highlighted in this analysis to fully explore disease pathogenesis.

Whilst the interactions identified require further independent validation, the unique experimental design using complementary capture baits (region and promoter captures) provides robust evidence of chromatin interactions. Additionally, validation with chromatin interactions identified by Rao *et al.* [19] further add to the confidence of the observed interactions. While Capture Hi-C offers much greater resolution for chromatin interactions than Hi-C, observed interactions are still limited by the restriction enzyme used and do not pinpoint the interactions to specific enhancers. As such further work will be necessary to confirm causal enhancers and how they affect gene regulation. The use of cell lines is a limitation of the study but the experimental requirement of high cell numbers for Capture Hi-C makes the use of primary cells challenging. However, it is essential that further experiments are performed in

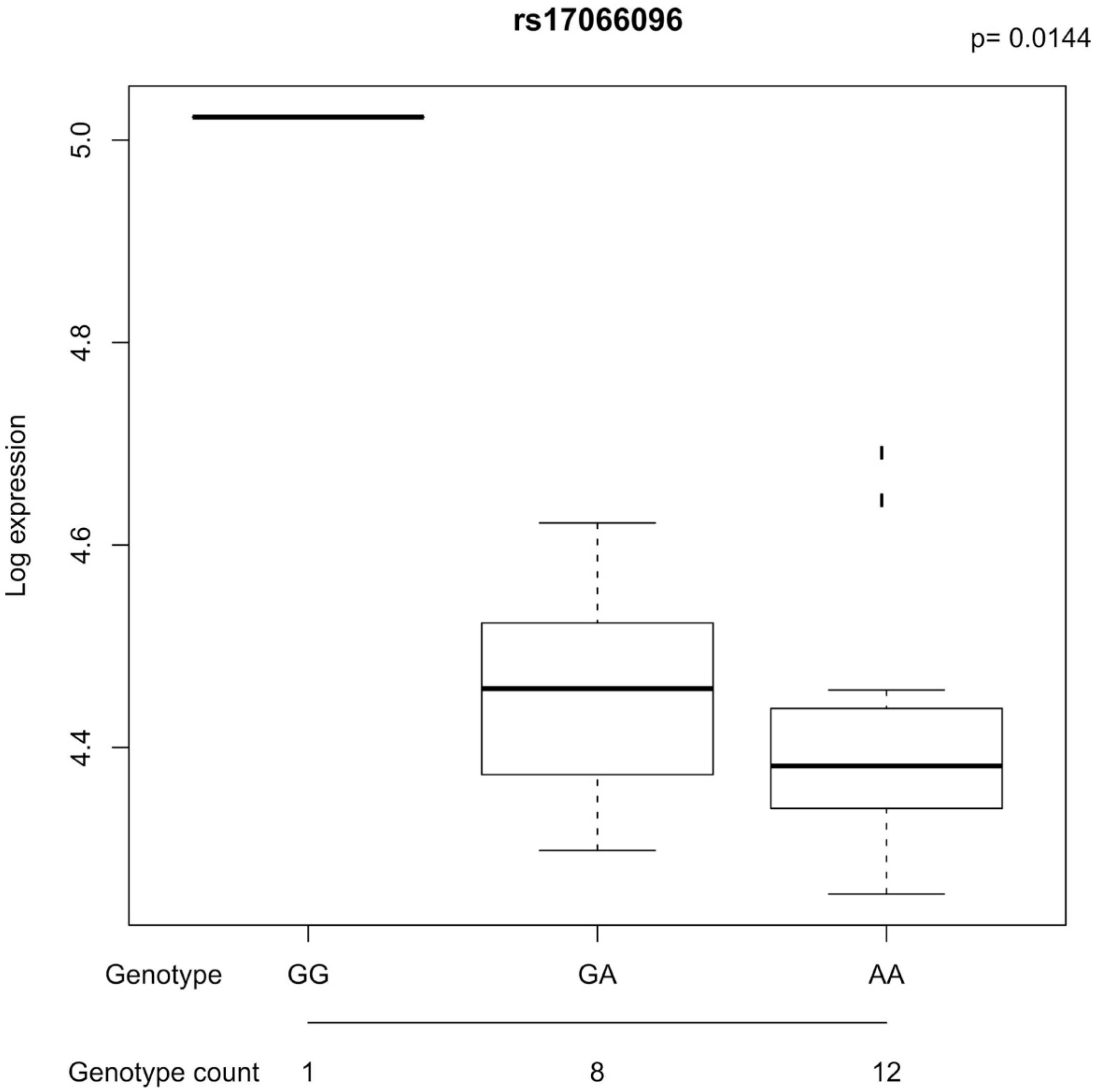


Fig 3. Increased expression of *IL20RA* in primary CD8+ T-cells from 21 healthy individuals carrying the G risk allele of rs17066096, *P* = 0.01. The three different genotypes for the SNPs are displayed on the X axis and gene expression levels on the Y axis. Error bars indicate standard deviation.

doi:10.1371/journal.pone.0166923.g003

primary cells to fully elucidate how chromatin interactions can effect gene regulation in MS. Despite these limitations of Capture Hi-C, it is clear that this technique is a powerful approach to link genes to their regulatory elements and this work has identified several candidate causal genes for MS. Additionally it has been proposed that by using genetic evidence to select drug targets, it could double the success rate in clinical development [41]. Since Capture Hi-C has

Table 3. Refined MS SNPs based on bioinformatics and Capture Hi-C data.

LD SNP	Index SNP	r^2	RegulomeDB Score	Promoter histone marks	Enhancer histone marks	DNAse	Proteins bound	Motifs changed	GRASP QTL hits	Selected eQTL hits	GENCODE genes	Gene Interactions*
rs11154801	rs11154801	1	1d	BLD	BLD	BLD		Nkx2	20 hits	46 hits	AHI1	*AHI1*, *ALDH8A1*, *HBS1L*, *MYB*
rs7759971	rs11154801	0.99	1f	BLD	BLD, STRM, SKIN				1 hit	46 hits	AHI1	*ALDH8A1*, *HBS1L*, *MYB*, *SGK1*
rs13197384	rs11154801	0.92	1a	24 tissues		53 tissues	42 bound proteins	13 altered motifs	1 hit	44 hits	AHI1	*BCLAF1*, *HBS1L*, *HMGB1P17*, *MTFR2*, *PDE7B*
rs7750586[†]	rs11154801	0.98	1f	GI, BLD	10 tissues	22 tissues	CTCF, RAD21, SMC3		2 hits	53 hits	LINC00271	
rs9399148	rs11154801	0.95	1f		8 tissues	LNG, GI, BLD	KAP1	GR	2 hits	52 hits	LINC00271	*MTFR2*, *PDE7B*
rs5880258	rs11154801	0.95			4 tissues	MUS, MUS		Pbx-1		21 hits	LINC00271	*MTFR2*, *PDE7B*
rs1322553	rs17066096	0.96	5	BLD	IPSC, BLD	BLD		4 altered motifs			44kb 3' of *IL22RA2*	*IFNGR1*, *IL22RA2*, *TNFAIP3*
rs17066063	rs17066096	0.96	3a	BLD	BRST, BLD, SKIN	7 tissues	6 bound proteins	Pou2f2, TCF4	1 hit		41kb 3' of *IL22RA2*	*IFNGR1*, *IL22RA2*, *TNFAIP3*
rs12214014	rs17066096	0.97	5		BLD			GR			36kb 3' of *IL22RA2*	*IL22RA2*, *TNFAIP3*
rs9321623	rs7769192	0.97	4	FAT, SKIN, MUS	15 tissues	7 tissues		Rad21			36kb 5' of RP11-95M15.1	*IL20RA*, *IL22RA2*
rs7769192	rs7769192	1	4		6 tissues	18 tissues	CTCF	HNF1, Irf, Pax-6			32kb 5' of RP11-95M15.1	*IFNGR1*, *IL20RA*, *IL22RA2*
rs11758213[‡]	rs7769192	1	2a		GI, BLD, LIV	39 tissues	CTCF, RAD21, SMC3	4 altered motifs			12kb 5' of RP11-95M15.1	
rs10607561	rs7769192	0.93	3a			BLD	NFKB, SETDB1	Foxp1, RREB-1			4.2kb 3' of RP11-95M15.1	*TNFAIP3*

r^2 has been rounded to 2 decimal places.

*—Only protein coding genes are shown for clarity.

[†] rs7750586 showed strong evidence of regulatory activity but was not involved in any interactions.

[‡] rs11758123 only interacts with lincRNAs. BLD, blood; BRST, breast; FAT, adipose tissue; GI, smooth muscle; IPSC, induced pluripotent stem cells; LIV, liver; LNG, lung; MUS, muscle; SKIN, skin; STRM, stromal connective.

doi:10.1371/journal.pone.0166923.t003

the potential to identify causal genes for genetic associations, it provides a way to enhance this further. This is exemplified by the identification of chromatin interactions between MS associations and the *IL20RA* and *IL22RA2* genes, showing a potential use of anti-IL20 therapy in MS, and highlights the potential of Capture Hi-C to provide novel therapeutic targets or drug repositioning to improve patient outcome.

Supporting Information

S1 Data. LD between MS associated SNPs and other disease associated SNPs in the 6q23 region. Disease abbreviations are as follows: CEL–Coeliac disease; CRO–Crohn's disease; MS–Multiple sclerosis; PBC–Primary biliary cirrhosis; PSO–Psoriasis; RA–Rheumatoid arthritis; SLE–Systemic lupus erythematosus; T1D –Type 1 diabetes; UC–Ulcerative colitis; SJO—Sjogren Syndrome.
(XLSX)

S2 Data. NRHV CD4 and CD8 eQTLs for MS SNPs within the 6q23 region.
(XLSX)

S3 Data. Full table of bioinformatics and Capture Hi-C data analysis. Refined SNPs are highlighted in green.
(XLSX)

S1 Fig. Interactions within the rs11154801 LD region. Tracks are labelled as follows: A–LD regions targeted in 'region' Capture Hi-C; B–Gene regions targeted in 'promoter' Capture Hi-C; C–GENCODE Genes V17; D–MS index SNPs; E–MS LD regions; F–Interactions observed in the GM12878 B-cell line and G–Interactions observed in the Jurkat T-cell line. All co-ordinates are based on GRCh37.
(TIF)

S2 Fig. Full overview of MS 6q23 Immunochip associated regions. Tracks are labelled as follows: A–LD regions targeted in 'region' Capture Hi-C; B–Gene regions targeted in 'promoter' Capture Hi-C; C–RefSeq genes (packed for clarity); D–MS index SNPs; E–MS LD regions; F–Interactions observed in the GM12878 B-cell line and G–Interactions observed in the Jurkat T-cell line. All co-ordinates are based on GRCh37.
(TIF)

Acknowledgments

The authors would like to acknowledge the Faculty of Life Sciences Genomics Facility and the assistance given by IT Services and the use of the Computational Shared Facility at The University of Manchester. We thank Dr Kathryn Steel for carrying out cell separation and gene expression profiling of healthy volunteers.

Author Contributions

Conceptualization: GO SE PF JW AB.

Formal analysis: PM SS GO JM.

Funding acquisition: AB JW PF SE GO.

Investigation: PM AM JM SS KD AY.

Methodology: PM GO SS AM PF SE.

Project administration: AB JW PF SE GO.

Resources: GO SE PF JW AB.

Supervision: AB JW PF SE GO.

Writing – original draft: PM GO.

Writing – review & editing: PM AM JM SS KD AY AB JW PF SE GO.

References

1. De Jager PL, Jia X, Wang J, de Bakker PIW, Ottoboni L, Aggarwal NT, et al. Meta-analysis of genome scans and replication identify CD6, IRF8 and TNFRSF1A as new multiple sclerosis susceptibility loci. Nat Genet. 2009; 41: 776–82. doi: 10.1038/ng.401 PMID: 19525953
2. Hafler DA, Compston A, Sawcer S, Lander ES, Daly MJ, De Jager PL, et al. Risk alleles for multiple sclerosis identified by a genomewide study. N Engl J Med. 2007; 357: 851–62. doi: 10.1056/NEJMoa073493 PMID: 17660530
3. Patsopoulos NA, Esposito F, Reischl J, Lehr S, Bauer D, Heubach J, et al. Genome-wide meta-analysis identifies novel multiple sclerosis susceptibility loci. Ann Neurol. 2011; 70: 897–912. doi: 10.1002/ana.22609 PMID: 22190364
4. Sawcer S, Hellenthal G, Pirinen M, Spencer CCA, Patsopoulos NA, Moutsianas L, et al. Genetic risk and a primary role for cell-mediated immune mechanisms in multiple sclerosis. Nature. 2011; 476: 214–9. doi: 10.1038/nature10251 PMID: 21833088
5. Hemminki K, Li X, Sundquist J, Hillert J, Sundquist K. Risk for multiple sclerosis in relatives and spouses of patients diagnosed with autoimmune and related conditions. Neurogenetics. 2009; 10: 5–11. doi: 10.1007/s10048-008-0156-y PMID: 18843511
6. Beecham AH, Patsopoulos NA, Xifara DK, Davis MF, Kemppinen A, Cotsapas C, et al. Analysis of immune-related loci identifies 48 new susceptibility variants for multiple sclerosis. Nat Genet. 2013; 45: 1353–60. doi: 10.1038/ng.2770 PMID: 24076602
7. Plenge RM, Cotsapas C, Davies L, Price AL, de Bakker PIW, Maller J, et al. Two independent alleles at 6q23 associated with risk of rheumatoid arthritis. Nat Genet. 2007; 39: 1477–82. doi: 10.1038/ng.2007.27 PMID: 17982456
8. Fung EYMG, Smyth DJ, Howson JMM, Cooper JD, Walker NM, Stevens H, et al. Analysis of 17 autoimmune disease-associated variants in type 1 diabetes identifies 6q23/TNFAIP3 as a susceptibility locus. Genes Immun. 2009; 10: 188–91. doi: 10.1038/gene.2008.99 PMID: 19110536
9. Graham RR, Cotsapas C, Davies L, Hackett R, Lessard CJ, Leon JM, et al. Genetic variants near TNFAIP3 on 6q23 are associated with systemic lupus erythematosus. Nat Genet. 2008; 40: 1059–61. doi: 10.1038/ng.200 PMID: 19165918
10. Nair RP, Duffin KC, Helms C, Ding J, Stuart PE, Goldgar D, et al. Genome-wide scan reveals association of psoriasis with IL-23 and NF-kappaB pathways. Nat Genet. 2009; 41: 199–204. doi: 10.1038/ng.311 PMID: 19169254
11. Dubois PCA, Trynka G, Franke L, Hunt KA, Romanos J, Curtotti A, et al. Multiple common variants for celiac disease influencing immune gene expression. Nat Genet. 2010; 42: 295–302. doi: 10.1038/ng.543 PMID: 20190752
12. Bowes J, Orozco G, Flynn E, Ho P, Brier R, Marzo-Ortega H, et al. Confirmation of TNIP1 and IL23A as susceptibility loci for psoriatic arthritis. Ann Rheum Dis. 2011; 70: 1641–4. doi: 10.1136/ard.2011.150102 PMID: 21623003
13. Jostins L, Ripke S, Weersma RK, Duerr RH, McGovern DP, Hui KY, et al. Host-microbe interactions have shaped the genetic architecture of inflammatory bowel disease. Nature. 2012; 491: 119–24. doi: 10.1038/nature11582 PMID: 23128233
14. Fairfax BP, Humburg P, Makino S, Naranbhai V, Wong D, Lau E, et al. Innate immune activity conditions the effect of regulatory variants upon monocyte gene expression. Science. 2014; 343: 1246949. doi: 10.1126/science.1246949 PMID: 24604202
15. Farh KK-H, Marson A, Zhu J, Kleinewietfeld M, Housley WJ, Beik S, et al. Genetic and epigenetic fine mapping of causal autoimmune disease variants. Nature. 2015; 518: 337–43. doi: 10.1038/nature13835 PMID: 25363779
16. Schoenfelder S, Clay I, Fraser P. The transcriptional interactome: gene expression in 3D. Curr Opin Genet Dev. 2010; 20: 127–33. doi: 10.1016/j.gde.2010.02.002 PMID: 20211559

17. Martin P, McGovern A, Orozco G, Duffus K, Yarwood A, Schoenfelder S, et al. Capture Hi-C reveals novel candidate genes and complex long-range interactions with related autoimmune risk loci. Nat Commun. 2015; 6: 10069. doi: 10.1038/ncomms10069 PMID: 26616563

18. Cairns J, Freire-Pritchett P, Wingett SW, Várnai C, Dimond A, Plagnol V, et al. CHiCAGO: robust detection of DNA looping interactions in Capture Hi-C data. Genome Biol. Cold Spring Harbor Labs Journals; 2016; 17: 127. doi: 10.1186/s13059-016-0992-2 PMID: 27306882

19. Rao SSP, Huntley MH, Durand NC, Stamenova EK, Bochkov ID, Robinson JT, et al. A 3D Map of the Human Genome at Kilobase Resolution Reveals Principles of Chromatin Looping. Cell. 2014; 159: 1665–80. doi: 10.1016/j.cell.2014.11.021 PMID: 25497547

20. Westra H-J, Peters MJ, Esko T, Yaghootkar H, Schurmann C, Kettunen J, et al. Systematic identification of trans eQTLs as putative drivers of known disease associations. Nat Genet. 2013; 45: 1238–43. doi: 10.1038/ng.2756 PMID: 24013639

21. Raj T, Rothamel K, Mostafavi S, Ye C, Lee MN, Replogle JM, et al. Polarization of the effects of autoimmune and neurodegenerative risk alleles in leukocytes. Science. 2014; 344: 519–23. doi: 10.1126/science.1249547 PMID: 24786080

22. Ward LD, Kellis M. HaploReg: a resource for exploring chromatin states, conservation, and regulatory motif alterations within sets of genetically linked variants. Nucleic Acids Res. 2012; 40: D930–4. doi: 10.1093/nar/gkr917 PMID: 22064851

23. Boyle AP, Hong EL, Hariharan M, Cheng Y, Schaub MA, Kasowski M, et al. Annotation of functional variation in personal genomes using RegulomeDB. Genome Res. 2012; 22: 1790–7. doi: 10.1101/gr.137323.112 PMID: 22955989

24. Valente EM, Brancati F, Silhavy JL, Castori M, Marsh SE, Barrano G, et al. AHI1 gene mutations cause specific forms of Joubert syndrome-related disorders. Ann Neurol. 2006; 59: 527–34. doi: 10.1002/ana.20749 PMID: 16453322

25. Dixon-Salazar T, Silhavy JL, Marsh SE, Louie CM, Scott LC, Gururaj A, et al. Mutations in the AHI1 gene, encoding jouberin, cause Joubert syndrome with cortical polymicrogyria. Am J Hum Genet. 2004; 75: 979–87. doi: 10.1086/425985 PMID: 15467982

26. Wu C, Yosef N, Thalhamer T, Zhu C, Xiao S, Kishi Y, et al. Induction of pathogenic TH17 cells by inducible salt-sensing kinase SGK1. Nature. 2013; 496: 513–7. doi: 10.1038/nature11984 PMID: 23467085

27. Kasof GM, Goyal L, White E. Btf, a novel death-promoting transcriptional repressor that interacts with Bcl-2-related proteins. Mol Cell Biol. 1999; 19: 4390–404. Available: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=104398&tool=pmcentrez&rendertype=abstract PMID: 10330179

28. Sharief MK, Matthews H, Noori MA. Expression ratios of the Bcl-2 family proteins and disease activity in multiple sclerosis. J Neuroimmunol. 2003; 134: 158–65. Available: http://www.ncbi.nlm.nih.gov/pubmed/12507784 PMID: 12507784

29. McPherson JP, Sarras H, Lemmers B, Tamblyn L, Migon E, Matysiak-Zablocki E, et al. Essential role for Bclaf1 in lung development and immune system function. Cell Death Differ. 2009; 16: 331–9. doi: 10.1038/cdd.2008.167 PMID: 19008920

30. Pollard KM, Cauvi DM, Toomey CB, Morris K V, Kono DH. Interferon-γ and systemic autoimmunity. Discov Med. 2013; 16: 123–31. Available: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3934799&tool=pmcentrez&rendertype=abstract PMID: 23998448

31. Ouyang W, Rutz S, Crellin NK, Valdez PA, Hymowitz SG. Regulation and functions of the IL-10 family of cytokines in inflammation and disease. Annu Rev Immunol. 2011; 29: 71–109. doi: 10.1146/annurev-immunol-031210-101312 PMID: 21166540

32. Gottlieb AB, Krueger JG, Sandberg Lundblad M, Göthberg M, Skolnick BE. First-In-Human, Phase 1, Randomized, Dose-Escalation Trial with Recombinant Anti-IL-20 Monoclonal Antibody in Patients with Psoriasis. PLoS One. 2015; 10: e0134703. doi: 10.1371/journal.pone.0134703 PMID: 26252485

33. Šenolt L, Leszczynski P, Dokoupilová E, Göthberg M, Valencia X, Hansen BB, et al. Efficacy and Safety of Anti-Interleukin-20 Monoclonal Antibody in Patients With Rheumatoid Arthritis: A Randomized Phase IIa Trial. Arthritis Rheumatol (Hoboken, NJ). 2015; 67: 1438–48. doi: 10.1002/art.39083 PMID: 25707477

34. Rutz S, Eidenschenk C, Ouyang W. IL-22, not simply a Th17 cytokine. Immunol Rev. 2013; 252: 116–32. doi: 10.1111/imr.12027 PMID: 23405899

35. Catrysse L, Vereecke L, Beyaert R, van Loo G. A20 in inflammation and autoimmunity. Trends Immunol. 2014; 35: 22–31. doi: 10.1016/j.it.2013.10.005 PMID: 24246475

36. Dong X, Li C, Chen Y, Ding G, Li Y. Human transcriptional interactome of chromatin contribute to gene co-expression. BMC Genomics. BioMed Central; 2010; 11: 704. doi: 10.1186/1471-2164-11-704 PMID: 21156067

37. Homouz D, Kudlicki AS. The 3D Organization of the Yeast Genome Correlates with Co-Expression and Reflects Functional Relations between Genes. Khodursky AB, editor. PLoS One. Public Library of Science; 2013; 8: e54699. doi: 10.1371/journal.pone.0054699 PMID: 23382942

38. Lan X, Witt H, Katsumura K, Ye Z, Wang Q, Bresnick EH, et al. Integration of Hi-C and ChIP-seq data reveals distinct types of chromatin linkages. Nucleic Acids Res. 2012; 40: 7690–7704. doi: 10.1093/nar/gks501 PMID: 22675074

39. Schoenfelder S, Sexton T, Chakalova L, Cope NF, Horton A, Andrews S, et al. Preferential associations between co-regulated genes reveal a transcriptional interactome in erythroid cells. Nat Genet. 2010; 42: 53–61. doi: 10.1038/ng.496 PMID: 20010836

40. Huseby ES, Huseby PG, Shah S, Smith R, Stadinski BD. Pathogenic CD8 T cells in multiple sclerosis and its experimental models. Front Immunol. 2012; 3: 64. doi: 10.3389/fimmu.2012.00064 PMID: 22566945

41. Tipney H, Painter JL, Shen J, Nicoletti P, Shen Y, Floratos A, et al. The support of human genetic evidence for approved drug indications. Nat Genet. Nature Research; 2015; 47: 1–7. doi: 10.1038/ng.3314 PMID: 26121088