Functional Characterisation of Rheumatoid Arthritis Risk Loci

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

2016

Amanda Jane McGovern

School of Biological Sciences

Contents

Contents

Final word count: 75,889

List of figures

List of Tables

List of Abbreviations

µg	Microgram
µl	Microlitre
µM	Micromolar
µm	Micrometre
^{o}C	Degrees centigrade
3C	Chromosome conformation capture
3C-qPCR	Chromosome conformation capture-quantitative polymerase chain reaction
3C-seq	Chromosome conformation capture sequencing
3-D	Three dimensional
4C	Chromosome conformation capture on chip/circular 3C
5C	Chromosome conformation capture carbon copy
5-mC	5-methyl cytosine
ACPA	Anti-cyclic citrullinated peptide antibodies
ACR	American College of Rheumatology
AFA	Adaptive focused acoustics
ARID5B	AT Rich Interactive Domain 5B
ATAC-seq	Assay for Transposase-Accessible Chromatin with high throughput sequencing
AZI2I	NF-Kappa-B-Activating Kinase-Associated Protein 1
BAC	Bacterial artificial chromosome
BACH2	BTB and CNC homology 1, basic leucine zipper transcription factor 2
BB	Binding buffer
BCL11A	B-cell lymphoma 11A
BCL3	B-cell lymphoma 3-encoded protein
bp	Base pair
BR	Broad-range
BSA	Bovine serum albumin
CBS	CTCF binding sites
CCL21	Chemokine ligand 21
CCND1	Cyclin D1
CCP	Cyclic citrullinated peptide
cDNA	Copy DNA
CeD	Coeliac disease
ChIA-PET	Chromatin interaction analysis with paired end tag sequencing
CHi-C	Capture Hi-C
ChIP	Chromatin immunoprecipitation
ChIP-Seq	ChIP-Sequencing
CLEC16A	C-Type Lectin Domain Family 16, Member A
cM	Centimorgan
cm^2	Centimetre squared
CNAP1	Chromosome condensation-related SMC-associated protein
CNC	Conserved non-coding
CO_2	Carbon dioxide
COG6	Component Of Oligomeric Golgi Complex 6
CRC	Colorectal cancer

CRISPR	Clustered regularly-interspaced short palindromic repeats
C_T	Threshold cycle
CTCF	CCCTC-binding factor
CTLA4	Cytotoxic T-Lymphocyte-Associated Protein 4
DAS	Disease activity score
dATP	Deoxyadenosine triphosphate
dCTP	Deoxycytidine triphosphate
DENND1B	DENN/MADD Domain Containing 1B
DEXI	Dexamethasone induced
DF	Dilution factor
dGTP	Deoxyguanosine triphosphate
DMARD	Disease modifying anti-rheumatic drug
DMEM	Dulbecco's Modified Eagle Medium
DNA	Deoxyribonucleic acid
DnaseI	Deoxyribonuclease
DnaseI-HS	DNaseI Hypersensitivity
DNMT	DNA methyltransferase
dNTP	Deoxynucleotide triphosphate
dsDNA	Double stranded DNA
ds-eQTL	DNase1 sensitivity eQTL
dTTP	Deoxythymidine triphosphate
EBF3	Early B-cell factor 3
EBV	Epstein-Barr Virus
EDTA	Ethylenediaminetetraacetic acid
EIF3H	Eukaryotic Translation Initiation Factor 3
EMSA	Electrophoretic mobility shift assay
EOMES	Eomesodermin
eQTL	Expression quantitative trait loci
ER-α	Oestrogen receptor alpha
eRNA	Enhancer RNA
EULAR	European League against Rheumatism
EWAS	Epigenome wide association study
FANTOM5	Functional annotation of the mammalian genome 5
FBS	Foetal bovine serum
FDR	False discovery rate
FISH	Fluorescent *in situ* hybridisation
FOXA1	Forkhead Box A1
FOXO1	Forkhead Box O1
FWD	Forward
GAPDH	Glyceraldehyde 3-phosphate dehydrogenase
Gb	Gigabytes
GWAS	Genome wide association study
HAT	Histone acetyltransferases
HBA	Haemoglobin A
HDAC	Histone deacetylase
HDM	Histone demethylases

HiCUP	Hi-C User Pipeline
HLA	Human leukocyte antigen
HMT	Histone methyltransferase
HPLC	High performance liquid chromatography
HRC	Human random control
HT1	Hybridisation buffer
HUVEC	Human umbilical vein endothelial cells
IBD	Inflammatory bowel disease
IBR	Indexed blocking reagent
IFNAR1	Interferon (Alpha, Beta And Omega) Receptor 1
IFN-β	Interferon beta
IFN-γ	Interferon gamma
IFNGR1	Interferon gamma receptor 1
IL-1	Interleukin-1
IL-2	Interleukin-2
IL-12	Interleukin-12
IL-17	Interleukin-17
IL20RA	Interleukin 20 receptor alpha
IL-6	Interleukin-6
IP	Immunoprecipitation
IRAK	Interleukin-1 Receptor-Associated Kinase
IRF-1	Interferon regulatory factor 1
IRX1	Iroquois Homeobox 1
JAK2	Janus kinase 2
JIA	Juvenile idiopathic arthritis
kb	Kilobases
KLF1	Kruppel-Like Factor 1 (Erythroid)
LB	Luria-Bertani (broth and agar)
LCL	B-lymphoblastoid cell line
LCR	Locus control region
LD	Linkage disequilibrium
lincRNA	Long intergenic non-coding RNA
LMA	Ligation mediated amplification
lncRNA	Long non-coding RNA
LR	Long range
LRI	Long range interaction
M	Molar
mAb	Monoclonal antibody
MAF	Minor allele frequency
MAPK	Mitogen-activated protein kinase
Mb	Megabase
mg	Milligram
miRNA	Micro RNA
ml	Millilitre
mM	Millimolar
MMP	Matrix metalloproteinase

MNase	Micrococcal nuclease
mRNA	Messenger RNA
MS	Multiple Sclerosis
MuTHER	Multiple Tissue Human Expression Resource
MW	Molecular weight
N-ChIP	Native chromatin immunoprecipitation
NCR	Negative control region
ncRNA	Non-coding RNA
NEB	New England Biolabs
NF-kB	Nuclear factor kappa-light-chain-enhancer of activated B-cells
ng	Nanogram
NGS	Next generation sequencing
NOD2	Nucleotide-Binding Oligomerisation Domain Containing 2
NTB	No Tween buffer
NTC	No template control
nM	Nanomolar
nm	Nanometre
NRHV	Arthritis Research UK National Repository for Healthy Volunteers
OLIG1	Oligodendrocyte transcription factor 1
OLIG2	Oligodendrocyte transcription factor 2
OLIG3	Oligodendrocyte transcription factor 3
pAb	Polyclonal antibody
PAGE	Polyacrylamide gel electrophoresis
PBMC	Peripheral blood mononuclear cells
PBS	Phosphate buffered saline
PcG	Polycomb group of proteins
PCR	Polymerase chain reaction
pg	Picogram
PHA	Phytohaemagluttinin
PICS	Probabilistic identification of causal SNPs
PLG	Phase-lock gel
pM	Picomolar
PolII	RNA polymerase II
PRKCQ	Protein kinase C theta
Ps	Psoriasis
PsA	Psoriatic arthritis
PTPN11	Protein Tyrosine Phosphatase, Non-Receptor Type 11
PTPN22	Protein tyrosine phosphatase, Non-Receptor Type 22
PTPRC	Protein Tyrosine Phosphatase, Receptor Type, C
PVT1	Pvt1 Oncogene (Non-Protein Coding)
PWAS	Proteome-wide association study
QC	Quality control
qPCR	Quantitative PCR
qRT-PCR	Quantitative reverse-transcription PCR
R_2	Correlation coefficient
RA	Rheumatoid arthritis

RAD23A	RAD23 Paralogue A
RAD51B	RAD51 Paralogue B
RANKL	Receptor activator of NF-κB ligand
RASF	Rheumatoid arthritis synovial fibroblast
RASGRP1	RAS Guanyl Releasing Protein 1 (Calcium And DAG-Regulated)
REV	Reverse
RF	Rheumatoid factor
RIF	Relative interaction frequency
Rn	Normalised reporter
RNA	Ribonucleic acid
RNaseA	Ribonuclease A
RNA-seq	RNA-sequencing
rpm	Revolutions per minute
RPMI	Roswell Park Memorial Institute
RT	Room temperature
RTKN2	Rhotekin 2
SDS	Sodium dodecyl sulphate
SLE	Systemic lupus erythromatosus
SNP	Single nucleotide polymorphism
SPRI	Solid phase reversible immobilisation
SORT1	Sortilin 1
sQTL	Splicing quantitative trait loci
SR	Short range
STAT4	Signal transducer and activator of transcription 4
St.Dev	Standard Deviation
T1D	Type 1 diabetes
T2C	Targeted chromatin capture
TAD	Topologically associated domain
TB	Tween buffer
TBE	Tris Borate EDTA buffer
TCC	Tethered conformation capture
TE	Tris-EDTA buffer
TF	Transcription factor
TFBS	Transcription factor binding site
TGFβ1	Transforming growth factor beta 1
TLE	Tris low-EDTA buffer
TLR	Toll-like receptor
Tm	Melting temperature
TNF	Tumour necrosis factor
TNFAIP3	Tumour necrosis factor alpha inducible protein 3
TOX3	TOX High Mobility Group Box Family Member 3
TRAF1	TNF receptor-associated factor 1
TRAF6	TNF receptor-associated factor 6
Tris-HCl	Tris(hydroxymethyl)aminomethane hydrochloride
TSS	Transcription start site
TXNDC5	Thioredoxin domain containing 5

TYK2	Tyrosine kinase 2
U	Unit
UCSC	University of California, Santa Cruz
VCAM-1	Vascular cell adhesion protein 1
VEGF	Vascular endothelial growth factor
WBI	Wash buffer 1
WBII	Wash buffer 2
WTCCC	Wellcome Trust Case Control Consortium
X-ChIP	Crosslinking ChIP
ZFP36L1	ZFP36 Ring Finger Protein-Like 1

The University of Manchester

Abstract of thesis submitted by Amanda Jane McGovern for the degree of Doctor of Philosophy entitled Functional Characterisation of Rheumatoid Arthritis Risk Loci in June 2016.

Rheumatoid arthritis (RA) is a complex autoimmune disease affecting approximately 1% of the population. Multiple factors contribute to the development of RA, with genetic factors accounting for around 60% of the disease risk. Over the last few years, genome-wide association studies (GWAS) have successfully been used to identify regions of the genome predisposing to complex disease. There are now 101 confirmed RA risk loci, but for the vast majority of these loci the causal gene and causal variant remain unidentified and therefore, their function in disease is unexplored. The majority of genetic variants, or single nucleotide polymorphisms (SNPs), associated with disease map to non-coding enhancer regions, which may regulate transcription through long-range interactions with their target genes.

The aims of this project were to identify the causal genes within an RA locus, pinpoint the causal variants and elucidate the mechanisms by which the variants modify gene function. Capture Hi-C (CHi-C) was carried out with the aim of identifying long range interactions between disease-associated SNPs and genes in four related autoimmune diseases. Many long-range interactions were identified which implicated novel candidate genes, interactions involving multiple genetic loci which had a common target, and interactions with loci which had previously been implicated in disease.

Complex interaction patterns were observed in many of the disease associated loci, particularly in the 6q23 locus which is associated with a number of autoimmune diseases and is the focus of the present thesis. Within the 6q23 locus, associated SNPs lie a large distance from any gene (>180kb) making it difficult to pinpoint the exact causal gene. Results from CHi-C and chromosome conformation capture (3C-qPCR) experiments indicated that restriction fragments containing disease associated intergenic SNPs could display genotype-specific interactions with genes associated with autoimmunity (*IL20RA* and *IFNGR1*). Interactions could also be detected with long non-coding RNAs (lncRNAs),

The lead SNP in the 6q23 region is in tight LD with eight other SNPs which are equally likely to be causal. Bioinformatics analysis suggested that the most plausible causal SNP in the 6q23 intergenic region was rs6927172, as it maps to an enhancer in both B-cells and T-cells, is in a DNaseI hypersensitivity cluster, shows transcription factor binding and is in a conserved region. Chromatin immunoprecipitation (ChIP) demonstrated binding of chromatin marks of active enhancers (H3K4me1 and H3K27ac) and the transcription factors BCL3 and NF-κB to the rs6927172 SNP target site in Jurkat T-cells and GM12878 B-cells, suggesting the risk allele could be associated with increased regulatory activity.

In conclusion, these results show that CHi-C can help identify novel GWAS causal genes with the potential to suggest novel therapeutic targets. For example *IL20RA* is already a target for a monoclonal antibody which has been shown to be effective in treating RA in clinical trials. This project has also provided compelling evidence that the autoimmune risk variant in the 6q23 locus, rs6927172, is within a complex gene regulatory region, involving multiple immune genes and regulatory elements, such as lncRNAs.

DECLARATION

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

COPYRIGHT STATEMENT

Acknowledgements

There are many people to thank for the help and advice they have given me whilst doing this PhD.

Firstly I thank my supervisors Dr Gisela Orozco and Dr Steve Eyre, and my advisor Dr Anne Hinks.

Special thanks go to Dr Stefan Schoenfelder and Dr Peter Fraser at the Babraham Institute, Cambridge for the training and advice in Capture Hi-C. I'd also like to thank Dr Caroline Ospelt at the University Hospital of Zurich for providing the synovial fibroblasts used in many of the 3C-qPCR assays, the Faculty of Life Sciences Genomics Facility who carried out the HiSeq runs, and Paul Martin for the bioinformatics support in analysing the Capture Hi-C data.

I am hugely grateful to Kate Duffus for being a great friend from the minute I started in this department and for the sanity breaks in meeting room C…

Finally, I could not have done this without the love and support of my amazing husband, family and friends who have all believed in me from the start and have kept me going. Thank you everyone – this is my magnificent octopus and I promise not to write anything else, ever!

1. Introduction

1.1. Rheumatoid arthritis

Rheumatoid arthritis (RA) is a complex, systemic auto-immune disorder caused by a combination of genetic and environmental factors. Approximately 1% of the population (Symmons 2005) are affected with RA, with more cases occurring in women in their fifth decade (reviewed in (Coenen *et al*. 2009) and (Kurko *et al*. 2013)). The main characteristics of RA are swollen joints, systemic inflammation and the presence of serum auto-antibodies such as rheumatoid factor (RF) and anti-cyclic citrullinated peptide antibodies (anti-CCP antibodies or ACPA) (De Rycke *et al.* 2004).

Evidence for a genetic role in RA was first reported in family studies in the early 20th century, when it was discovered that individuals with first-degree relatives with RA had 2-4 times the risk of developing the disease compared to those without relatives with RA (Terao *et al.* 2016). Twin studies have added further evidence for a genetic role in RA (Macgregor *et al.* 2000; Silman *et al.* 1993) with monozygotic twins having approximately 15% phenotypic concordance compared to 3% in dizygotic twins.

The heritability of a disease is an estimate of how much of the likelihood of developing a disease is due to genetics. From various twin studies in mostly European populations, the heritability of RA has been estimated to be between 40-60% (Terao *et al.* 2016). Approximately 12% of the heritability of RA can be explained by associations with the MHC (major histocompatibility complex, discussed later) (Terao *et al.* 2016). Associations with all other genetic factors accounts for approximately 23% of RA heritability (Okada *et al.* 2014; Stahl *et al.* 2010), making a combined heritability of 35%.

Environmental factors contributing to RA include smoking (Kallberg *et al.* 2011; Klareskog *et al.* 2011; Silman 1993; Symmons *et al.* 1997), exposure to silica (Klockars *et al.* 1987; Stolt *et al.* 2005; Turner *et al.* 2000) and mineral oil (Sverdrup *et al*. 2005). The ratio of females to males with RA is 3:1 (Viatte *et al*. 2013), therefore sex hormones could be playing a role in RA susceptibility (Cutolo *et al.* 2006; Cutolo 2007; Luckey *et al.* 2012; Masi *et al.* 2006; Viatte *et al.* 2013). Evidence for a possible epigenetic role in sex susceptibility is shown by an association at the *IRAK1* locus (encoding Interleukin-1 Receptor-Associated Kinase 1), a gene which escapes X-inactivation (Carrel *et al.* 2005).

Epigenetic mechanisms are changes in deoxyribonucleic acid (DNA) that do not change the DNA sequence and are thought to play a role in RA pathogenesis. The role of epigenetics in autoimmunity is discussed further in section 1.3.5 and there are many recent reviews (Bottini *et al.* 2013; Lu 2013; Viatte *et al.* 2013).

There is also evidence that the microbiome, a complex population of bacteria, viruses and fungi inhabiting the human body, could be an important factor in the development of RA and other autoimmune diseases. The role of the microbiome is reviewed in (Belkaid *et al.* 2014) and (Yeoh *et al.* 2013). RA patients have been shown to have a high prevalence of periodontitis and tooth loss, implicating the oral microbiota in RA pathogenesis (Loyola-Rodriguez *et al.* 2010). The

intestinal microbiota is also thought to play a role in RA but the exact mechanisms are unknown (Edwards 2008). A murine model of RA showed that there was a reduction in disease activity in the absence of the microbiome resulting in a reduced Th17 response, an important mediator of RA pathogenesis (Wu *et al*. 2010). However, the exact mechanisms of how the microbiome contributes to RA are unknown (Scher *et al.* 2011).

1.1.1. RA pathogenesis

Cells of the immune system and synovium are a key factor in driving a complex process, causing the prolonged inflammation which is characteristic of RA, and ultimately resulting in the destruction of synovial joints. Cartilage destruction is caused by TNF, IL-1 and IL-6 stimulating the release of matrix metalloproteinases (MMPs) (Sabeh *et al*. 2010) and cathepsin k (Hou *et al.* 2001) resulting in the degradation of type II collagen and aggrecan fibres. Destruction of bone is due mainly through the action of osteoclasts which are stimulated by expression of the receptor activator of NF-κB ligand (RANKL). Further bone destruction occurs through the direct action of T-cells on osteoclasts and the presence of synoviocytes in the inflamed regions (Schett *et al.* 2012; Smith *et al.* 2002; Udagawa *et al.* 2002).

1.1.1.1. Cells of the immune system

The characteristic inflammation of RA is mainly driven by the cells of the adaptive immune system (Klareskog *et al.* 2009; McInnes *et al.* 2011). T-lymphocytes that are activated during the inflammatory response in turn trigger the activation of macrophages which release cytokines such as interleukins -1 and -6 (IL-1 and IL-6) and tumour necrosis factor (TNF), all of which drive a pro-inflammatory cascade (Choy 2008; Choy 2012; McInnes *et al.* 2007). Activated B-cells produce antibodies and antigen presenting cells, further contributing to the immune response (Zhang *et al.* 2001).

Recently CD4+ T-cells, especially CD4+ memory and regulatory T-cells (T_{REG}), have been identified as a critical cell type in RA (Diogo *et al.* 2014; Trynka *et al.* 2013) using the known genetic associations and bioinformatics analysis tools. Signalling pathways implicated in RA pathogenesis include those regulating T-cell activation, JAK-STAT signalling pathway, and NF-κB (nuclear factor kappa-light-chain-enhancer of activated B-cells) signalling. NF-κB signalling regulates genes involved in immunity, inflammation and, cell survival eg. *CD40, REL, TNFAIP3* (tumour necrosis factor alpha inducible protein 1/A20), *TRAF1* (TNF receptor-associated factor 1), *CCL21* (chemokine ligand 21), *PRKCQ* (protein kinase C theta), and *TRAF6* (TNF receptor-associated factor 6) and is induced by cytokines TNF-α and IL-1β and by CD40 engagement. The JAK-STAT pathway is the principle response to cytokines eg IL-6.

The role of B-cells in RA is less defined, but recently it has been suggested that activated human B-cells interact with synovial fibroblasts, inducing conversion of normal fibroblasts into inflammatory fibroblasts with an aggressive phenotype, resulting in prolonging inflammation. It has also recently

been shown that B-cell derived cytokines such as TNF and IL-1 are mediators of synovial fibroblast activation (Storch *et al.* 2016).

1.1.1.2. Synovial fibroblasts

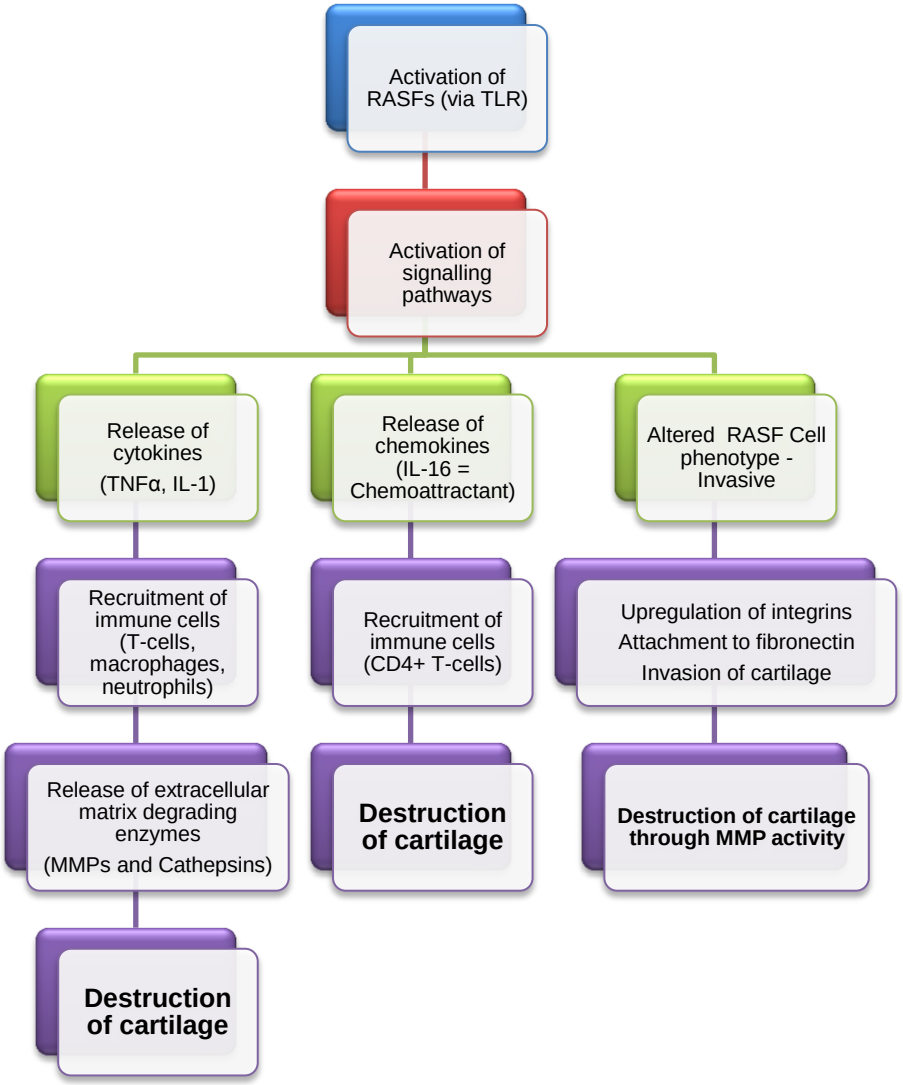
Another key cell-type involved in RA pathogenesis are synovial fibroblasts which have been known for a long time to have a major role in the initiation and progression of RA (Franz *et al.* 1998). The role of synovial fibroblasts in RA pathogenesis is complicated, and has been comprehensively reviewed (Huber *et al.* 2006). It is thought that cytokines derived from synovial fibroblasts and other sources, such as interleukin-6 (IL-6), help immune cells expand into the synovium in an antigen-independent manner (Huber *et al.* 2006).

RA synovial fibroblasts (RASFs) have been shown to behave and appear very differently to normal fibroblasts. When activated, RASFs become rounded in appearance, with a large nucleus and prominent nuclei which is indicative of active RNA metabolism. Activated RASFs show an invasive phenotype compared to normal fibroblasts whereby RASFs can be cultured without evidence of contact inhibition and readily attach to articular cartilage and invade the extracellular matrix (Huber *et al.* 2006; Lafyatis *et al.* 1989; Muller-Ladner *et al.* 1995). Proliferation and migration of RASFs has been shown to be mediated by TGFβ1 (transforming growth factor beta 1) (Bira *et al.* 2005).

RASFs contribute to RA pathogenesis and the immune response in a number of ways, summarised in Figure 1. Upon activation, RASFs express toll-like receptors (TLR), particularly TLR-2 and TLR-3, on the cell surface which are key receptors for the innate immune system (Brentano *et al.* 2009; Hu *et al.* 2014). Interactions with TLRs result in the production of pro-inflammatory cytokines, MMPs, vascularisation factors such as VEGF (vascular endothelial growth factor), and activates signalling pathways such as NF-κB, MAPK (mitogen activated protein kinases) and IRF3 (interferon regulatory factor 3). TLRs contribute to activation of Th1 and Th17 cells, resulting in further release of pro-inflammatory cytokines interferon gamma (IFN-γ) and IL-17 (Hu *et al.* 2014). TLR-2 has been shown to induce migration and invasion of RASFs, suggesting a potential therapeutic target (McGarry *et al.* 2015).

Attachment of RASFSs to the extracellular matrix is mediated through integrins through regions rich in fibronectin, type II collagen and glycosaminoglycans. Adhesion molecules such as VCAM-1 (Vascular cell adhesion protein 1) activate signalling cascades involved in the cell-cycle which results in overexpression of MMPs and key regulatory genes such as *cMyc* (Huber *et al*. 2006).

Figure 1: The role of synovial fibroblasts in RA pathogenesis



Common outcomes of prolonged inflammation in RA include fatigue due to the action of cytokines IL-1 and IL-6 on prostaglandin signalling pathways in brain endothelial cells (Klareskog *et al.* 2009).

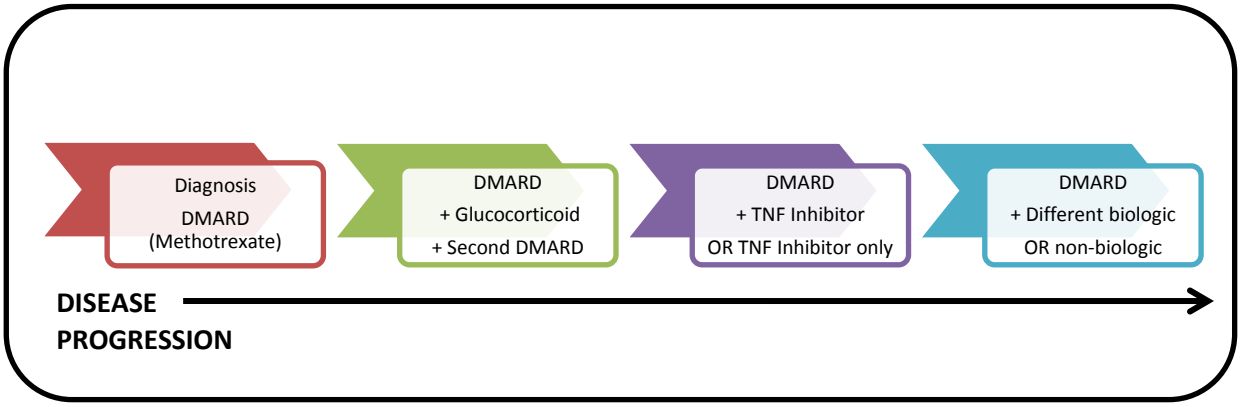
RA patients have an increased risk of cardiovascular disease related to prolonged inflammation (Gabriel *et al.* 2012). There is also a well-documented increase in the risk of lymphoma associated with RA and other auto-immune disorders (Starkebaum 2007).

1.1.2. Treatment of RA

Treatment strategies in RA aim to reduce inflammation early in the disease to minimise joint damage. Classification criteria have been developed to assist in choosing the most effective therapy according to disease progression. The ACR (American College of Rheumatology) criteria measures relative changes in RA symptoms (Singh *et al*. 2012), whereas the Disease Activity Score (DAS) is an absolute measure of disease activity (Fransen *et al.* 2006). The European League against Rheumatism (EULAR) criteria uses a combination of the ACR and DAS criteria (Smolen *et al.* 2010) and assesses joint involvement, the presence of RF and anti-CCP auto-antibodies (ACPA), markers of inflammation and how long the symptoms have been present.

Early treatment is usually carried out using disease modifying anti-rheumatic drugs (DMARDs) such as methotrexate, sulfasalazine, and glucocorticoids which control inflammation and reduce the development of joint erosions (Figure 2). If there is a poor response to early treatment, targeted therapies can be prescribed (biologics) which aim to control specific immune system pathways involved in RA pathogenesis (Choy *et al*. 2013).

Figure 2: Disease progression and treatment in RA



Examples of biological therapies used in RA include TNF blocking agents such as etanercept, which acts by partly neutralising circulating TNF. Other inflammatory molecules can be targeted by existing biological therapy such as IL-1 and IL-6 with anakinra and tocilizumab respectively. Certain cell populations can be targeted, for example, abatacept inhibits T-cell activation and rituximab depletes B-cell populations (Choy *et al.* 2013; Geiler *et al.* 2011; Smolen *et al.* 2010).

1.2. The genetic basis of RA

The identification of causal variants and causal genes in RA is of utmost importance in order to develop improvements in diagnosis and find novel therapeutic targets. The most common genetic variation associated with disease is the single nucleotide polymorphism (SNP) which is a single base change in the DNA sequence. SNPs can result in changes in the amino acid sequence of a protein (non-synonymous), which in turn can alter the protein function. Synonymous SNPs alter the DNA sequence but the amino acid sequence is unaffected. However, most of the SNP associations seen with RA, and all complex diseases, are with SNPs that reside outside a traditional protein coding region. These associated SNPs are likely to function by regulating the amount of protein, rather than changing the protein itself.

1.2.1. Identification of RA risk loci

The first loci associated with RA were identified through candidate gene studies and linkage analysis but in recent years, the development of genome wide association studies (GWAS) has been instrumental in identifying genetic association. Through candidate gene studies and linkage analysis the most significantly associated RA loci were identified as *HLA-DRB1* (human leukocyte antigen) and *PTPN22* (protein tyrosine phosphatase non-receptor type 22) which account for approximately 40% of the genetic component of RA (Morgan *et al.* 2010).

1.2.1.1. The HLA locus

The *HLA-DR* locus was first identified as being associated with RA in a population serology study by Stastny in 1979 (Stastny *et al.* 1979). The *HLA-DR* locus is part of the major histocompatibility complex (MHC) Class II gene family on chromosome 6p21 which encode transmembrane glycoproteins containing heterodimeric α and β chains, displayed on the surface of antigen presenting cells (Deighton *et al*. 1989). Molecular typing studies in 1987 led to the discovery of the 'shared epitope' (Gregersen *et al.* 1987), a conserved amino acid sequence (QXRAA) in the third hypervariable region of the HLA-DR β chain at positions 70-74 that is significantly associated with the development of anti-CCP antibodies (ACPA) and the development of ACPA-positive RA. The mechanism proposed that differences in alleles could alter antigen presentation or the representative T-cell population. A recent study (Raychaudhuri *et al*. 2012) identified a region outside of the shared epitope that is strongly associated with RA, localising to amino acid 11 in the HLA-DR β chain, and better explains the association at this complicated locus. Additional association was also confirmed at amino acid positions 71 and 74 within the shared epitope.

1.2.1.2. Non-HLA loci

The *PTPN22* locus has been associated with many auto-immune disorders (Fousteri *et al*. 2013) and was the first locus outside of the MHC to be robustly associated with RA. *PTPN22* encodes protein tyrosine phosphatase non-receptor 22 which is a key regulator of T-cell receptor signalling. Candidate gene studies linked *PTPN22* to type 1 diabetes (T1D) (Bottini *et al.* 2004) by associating the minor allele of a non-synonymous SNP (rs2476601) that resulted in an amino acid change at

position 620 (R620W). Linkage analysis and association studies confirmed association with RA in many populations (Begovich *et al.* 2004; Stanford *et al.* 2014). Recent studies using mouse models have given insight into the role of *PTPN22* in autoimmunity, showing that the R619W mutation in mice results in spontaneous autoimmunity (Zheng *et al.* 2014).

One other significant locus, with more modest association, has also been identified as being involved in RA susceptibility in candidate gene and linkage analysis studies (Ji *et al.* 2010; Orozco *et al.* 2008; Remmers *et al.* 2007). *STAT4* (signal transducer and activator of transcription 4) is a member of a transcription factor family involved in IFN-γ cytokine receptor signalling activated by IL-12 signalling through *JAK2* (janus kinase 2) in T-cells (Kurko *et al.* 2013).

1.2.2. Genome wide association studies (GWAS)

The completion of the human genome project (Venter *et al*. 2001) and the International HapMap Project in 2003 (HapMap 2003) which enabled the use of genotyping arrays to capture variations on a genome-wide scale have revolutionised the study of complex genetic diseases.

GWAS are based on genetic association, determined if variants are seen more or less frequently in disease cohorts compared to cohorts of unaffected, healthy populations. Statistical tests are used to correct for large numbers of tests that could lead to false positive results. For an association to have genome wide significance, the p value must be $\leq 5x10^{-8}$ and be independently replicated in another study (Ricano-Ponce *et al.* 2013). SNPs are used as markers of genetic susceptibility, with common SNPs (minor allele frequency (MAF) >1%) underlying many common diseases. GWAS genotyping arrays contain 'tag SNPs' which act as markers for candidate causal variants, many in non-coding regulatory regions that could possibly have a functional effect. As GWAS progress it is necessary to analyse data from many more individuals (Edwards *et al.* 2013), usually from different populations, in meta-analyses to generate the power needed to find more modestly associated and/or rarer associated variants. A database of GWAS studies at http://www.ebi.ac.uk/gwas contains 37 RA GWAS included up to March 2016 (Welter *et al.* 2014).

The Wellcome Trust Case Control Consortium (WTCCC) carried out the first major GWAS in 2007 (WTCCC 2007). The WTCCC consists of 50 research groups from across the UK which between them analysed 15,000 samples across 7 common complex diseases – bipolar disorder, coronary artery disease, Crohn's disease, rheumatoid arthritis, hypertension, type 1 diabetes and type 2 diabetes. The groups analysed 2000 UK cases for each disease and 3000 shared controls using an Affymetrix gene chip 500K SNP array. The WTCCC GWAS identified 24 independently associated regions across the 7 diseases, with previously identified RA regions HLA and *PTPN22* showing association to RA at genome-wide significance ($P<5x10^{-8}$). Nine other loci showed association at $P=1x10^{-5} – 5x10^{-7}$ and modest association of 49 SNPs was also detected at $P=1x10^{-4} – 5x10^{-5}$.

To follow up SNPs identified in GWAS, validation studies are required to determine true associations from false positives. For example, the 6q23 region was identified in the WTCCC

GWAS as having modest association with RA along with 8 other loci. An associated SNP in this UK based study, rs6920220, along with a second independently associated SNP from a United States cohort study, map to an intergenic region between the *TNFAIP3* (A20) and *OLIG3* (encoding oligodendrocyte transcription factor 3) genes (Plenge *et al.* 2007; Thomson *et al.* 2007). These SNPs were both validated in well powered replication studies, indicating they are truly associated with disease susceptibility. *TNFAIP3* itself has also been linked to systemic lupus erythematosus (SLE) (Graham *et al.* 2008) and is a strong candidate gene for RA, being strongly involved in inflammation (Vereecke *et al.* 2011). A20 is a potent anti-inflammatory molecule and is a negative regulator of NF-κB responses to TNF-α, toll-like receptor (TLR) and NOD2 (nucleotide-binding oligomerisation domain containing 2) signalling (Vereecke *et al.* 2011).

In 2010, a GWAS meta-analysis by Stahl *et al* (Stahl *et al.* 2010) studied 5,539 RA cases and 20,169 controls across European populations, followed by a replication study in 6,768 RA cases and 8,806 controls. Using this approach, 7 new RA loci near genes involved in immunity were identified at genome-wide significance ($P<5x10^{-8}$), taking the number of RA associated loci up to 31 in individuals of European ancestry.

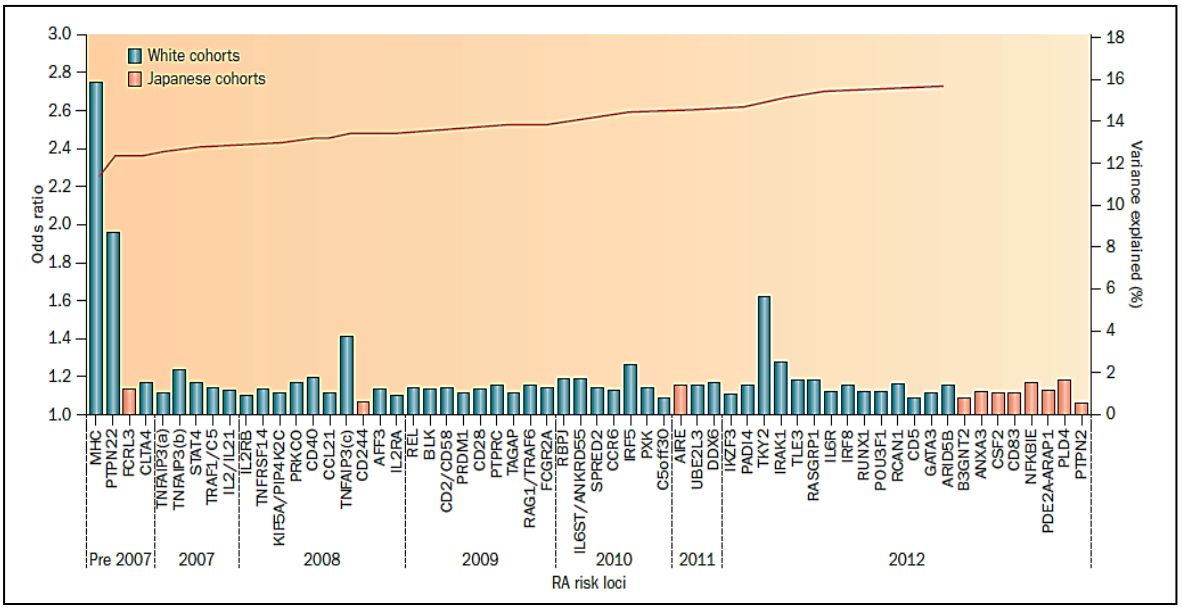
An extensive genetic fine-mapping study in 2012 (Eyre *et al.* 2012) carried out genotyping of 11,475 cases and 15,870 controls using a custom SNP array (Immunochip, discussed below) to analyse 130,000 SNPs. Analysis of 186 genetic loci discovered and fine-mapped 14 novel RA risk loci, along with fine-mapping 19 previously identified loci. The number of RA loci in individuals of European ancestry after this study was expanded to 46 (Figure 3 – from Viatte *et al* 2013).

To validate the findings from the dense-mapping study (Eyre *et al.* 2012), a validation of suggestively implicated variants ($P<1x10^{-5}$) was performed in 2013 (McAllister *et al*. 2013). From this study two SNPs were confirmed as being associated with RA. One SNP, rs72928038, mapped to an intron of *BACH2* (BTB and CNC homology 1, basic leucine zipper transcription factor 2), and the second, rs911263, to an intron of *RAD51B* (RAD51 paralogue B). The identification of these genes implicated two pathways in RA pathogenesis responsible for B-cell differentiation and DNA repair and brought the number of RA loci to 48.

Recently, the RACI consortium (Okada *et al*. 2014) performed a GWAS meta-analysis on more than 100,000 European and Asian samples. Samples were combined from different populations to make a trans-ethnic study, increasing the statistical power to detect novel loci. Data from 22 GWAS were combined to include 29,880 RA cases and 73,758 controls. One million SNPs from the 1000 Genomes Project were evaluated (Abecasis *et al*. 2012) which identified a further 42 novel RA risk loci at genome-wide significance, taking the number of RA risk loci to 101 and 98 candidate genes. Many of the candidate genes identified in this study overlapped with drug target genes, highlighting the importance of GWAS in drug discovery.

Increasing the number of samples in GWAS studies adds statistical power and the ability to potentially identify new disease susceptibility loci. A recent example of an extended GWAS identified a novel RA susceptibility locus at 22q12, confirmed most of the known associations with RA and increased the strength of association in some of the loci (Orozco *et al.* 2014).

Figure 3: RA loci identified through GWAS up to 2012



From Viatte *et al* 2013

1.3. The post-GWAS genetic landscape

GWAS have been tremendously successful in identifying SNPs associated with complex diseases, but the vast majority of causal SNPs and causal genes have not been identified. GWAS arrays are designed such that each genotyped SNP tags a large number of un-typed SNPs. This has the advantage of providing information genome-wide for many SNPs strongly correlated with the genotyped SNP through linkage disequilibrium (LD), although this correlation to many other SNPs has the disadvantage of making it difficult to identify the actual causal gene or variant. Fine-mapping studies can enable further refinement of the genetic signal, localising the associated variants and identifying SNPs which could be disease-causing. Once the risk loci have been fine-mapped, functional studies are still necessary as part of post-GWAS investigations in order to confirm causal SNPs and the genes on which they act. This will address the question of how genetic variation affects gene function and elucidate the molecular mechanism underlying the phenotype (Edwards *et al.* 2013; Freedman *et al.* 2011; Viatte *et al.* 2013).

1.3.1. Fine-mapping of genetic risk loci

Fine-mapping studies use publicly available data from sources such as the 1000 genomes project (1KG) (Abecasis *et al.* 2012) to design dense custom genotyping arrays, in an attempt to localise the true causal variants in a risk locus. The Immunochip project (Trynka *et al*. 2011) was the first major fine-mapping study of autoimmune disease associated regions. Many research groups collaborated to design a custom Illumina SNP microarray (Immunochip) containing ~200,000 SNPs across 186 risk loci which had previously been identified by GWAS. The Immunochip study identified 14 novel RA risk loci and refined the location of 19 previously associated RA risk loci.

A previous exemplar to this study showing how fine-mapping can be used to localise association and identify additional variants is from a study on the *TNFAIP3-OLIG3* intergenic region (Orozco *et al*. 2009). Fine mapping of the 6q23 locus found three independently associated SNPs - rs6920220, rs5029937 and rs13207033 and was the first to show that associated variants identified in GWAS loci, with only modest effect sizes, could produce a significantly greater effect once other risk variants within the region were considered.

1.3.2. The effects of SNPs on protein function

As previously mentioned, coding SNPs can be synonymous (do not affect amino acid sequence) or non-synonymous (affect amino acid sequence). Non-synonymous SNPs can have many effects on proteins such as truncation through the addition of a premature stop codon or alteration of folding resulting in loss of function. An example of a non-synonymous SNP affecting protein function is in *PTPN22 (*Begovich *et al.* 2004*)*. The rs2476601 SNP changes a C to T in the DNA sequence which causes an amino acid change from arginine to tryptophan at position 620 (R620W) resulting in a structural change within the PTPN22 polypeptide chain in a potential Src binding site (Gregersen 2005).

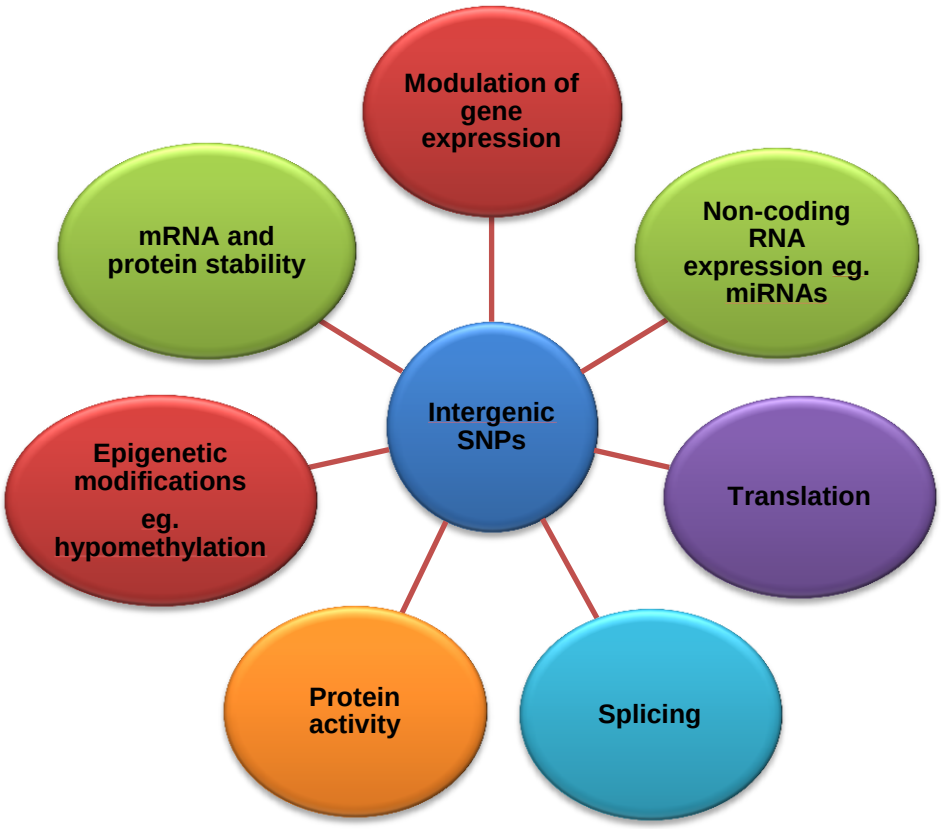
Other genes associated with RA are affected by non-synonymous SNPs, for example, rs8192284 at the *IL6R* locus (Eyre *et al.* 2012; Lamas *et al.* 2010) is associated with higher serum levels of soluble IL6R (sIL-6R). The tyrosine kinase 2 (*TYK2*) polymorphism rs34536443 causes an amino acid substitution from Proline to Alanine within the kinase domain of the TYK2 protein, and has been associated with RA, psoriatic arthritis (PsA) and multiple sclerosis (MS) (Ban *et al.* 2009; Bowes *et al.* 2015; Couturier *et al.* 2011; Okada *et al.* 2014).

The importance of non-synonymous SNPs is especially relevant if proteins such as transcription factors are affected, which could affect expression of many other downstream genes.

1.3.3. Investigating the role of intergenic variants

However, the majority of GWAS hits (>90%) (Farh *et al.* 2015) are in intergenic, regulatory regions (Maurano *et al.* 2012; Visel *et al.* 2009). A large percentage (~60%) map to immune cell enhancers and 10-20% directly alter a transcription factor binding motif (Farh *et al.* 2015). Therefore, their impact on disease could be the result of a potential ability to differentially regulate gene expression or through other mechanisms (Figure 4). The use of publicly available human genome (ENCODE) and epigenome (Epigenomics Roadmap, Blueprint) datasets showing transcription factor binding sites and chromatin states in many cell types and stimulatory conditions across the genome (Bernstein *et al.* 2010; ENCODE Project 2012; Martens *et al.* 2013) can aid in the annotation of these intergenic associated variants.

Figure 4: Potential mechanisms of action of intergenic SNPs



The first stage of gene expression is the transcription of genomic DNA into RNA (ribonucleic acid) by RNA polymerase II (PolII). Specific sequences of DNA known as transcription start sites (TSS), or core promoters, assemble the core PolII machinery. Transcription factors are DNA-binding proteins which bind accessory proteins and short DNA motifs which can be promoters, enhancers, silencers or insulators and affect how much a gene is transcribed (Maston *et al*. 2006). There are far fewer transcription factors than transcribed genes, therefore a complex combination of

transcription factors and regulatory elements such as enhancers is required for gene expression (Maston *et al.* 2006)

Promoters are located ~500bp upstream from the transcription start site for a gene and contain binding sites for activators (Maston *et al.* 2006). Enhancers are responsible for initiating gene transcription when bound by transcription factors (Pennacchio *et al.* 2013). Regions of DNA containing active enhancers and that are transcriptionally active are free of nucleosomes, have accessible DNA, have binding sites for transcription factors (TFBS) and contain post-translational modifications such as H3K4me1 and H3K27ac. Enhancers can lie a considerable linear distance away from their target gene, however, the 3-D folding of chromatin brings enhancers into close proximity to promoters. Transcription is initiated through interaction with activator proteins that bind to the mediator complex which recruits general transcription factors and RNA polymerase II (Maston *et al.* 2006). Silencers contain binding sites for repressor transcription factors. Insulators, such as CCCTF binding factor (CTCF), act by blocking enhancer-promoter interactions, or through the physical separation of chromatin into domains which prevents an enhancer contacting its promoter (Maston *et al.* 2006).

There are >100,000 enhancers, the majority of which show some cell type specificity (ENCODE Project 2012; Heintzman *et al.* 2007). Activation requires recognition sequences which allow the binding of transcription factors and accessory factors necessary to initiate transcription (Rosenfeld *et al.* 2006). Corradin *et al* (Corradin *et al.* 2014) studied six autoimmune diseases and showed that multiple variants within an LD block affected multiple enhancers, altering gene expression (multiple enhancer variant hypothesis). This has also been shown in prostate cancer risk loci (Zhang *et al.* 2012c), therefore the relationship between disease associated variants and enhancers is likely to be both cell specific and complicated.

There are ~2000 transcription factors in the human genome (Maston *et al.* 2006) and ~200-300 are expressed per cell type (Vaquerizas *et al.* 2009). Cell type specific interactions are enriched for enhancer-promoter interactions and are correlated with differential gene expression (Heidari *et al.* 2014). It has been shown that active genes and regulatory factors can arrange into genomic hotspots known as transcription factories, which enables co-ordinated control of transcription (Schoenfelder *et al.* 2010b). Interestingly, transcription factors which are critical for important biological processes have been shown to cluster together (Hnisz *et al.* 2013; Whyte *et al.* 2013) to control the transcription of the key genes. Super-enhancers are defined as a large cluster of enhancers, occupied by key transcription factors and the Mediator coactivator, that drive the expression of genes controlling cell identity (Hnisz *et al.* 2013; Pott *et al.* 2015). Liu *et al* investigated oestrogen regulated enhancers and found that a group of six transcription factors associated with oestrogen receptor alpha (ERα) that they named the mega-trans complex (Liu *et al.* 2014).

1.3.4. Expression quantitative trait loci - eQTLs

Genomic regions containing variants that have an effect on gene expression are known as expression quantitative trait loci (eQTLs). Variants can affect the expression of nearby genes within 1Mb (*cis*-eQTLs) or far away genes (*trans*-eQTLs) and are tissue and cell type specific (Edwards *et al.* 2013; Fu *et al.* 2012; Maurano *et al.* 2012; Nica *et al.* 2010).

Cis-eQTLs have been studied extensively using lymphoblastoid cell lines (LCLs) from HapMap populations (Cheung *et al.* 2005; Dimas *et al.* 2009; Stranger *et al.* 2005; Stranger *et al.* 2012). Stranger *et al* (Stranger *et al.* 2012) studied cis-eQTLs in 8 HapMap populations in ~20,000 gene expression phenotypes and found that ~20% of genes had a common *cis*-eQTL in at least 1 population and that clusters of *cis*-eQTLs were situated around transcription start sites. An example of a *cis*-eQTL was found when a SNP in a conserved region of 8q23.3 (rs16888589) was associated with allele-specific increase in expression of *EIF3H* (Eukaryotic Translation Initiation Factor 3) in colorectal cancer (CRC) cell lines (Pittman *et al*. 2010). *Trans*-eQTLs have more subtle effects and are tissue-specific (Majewski *et al.* 2011).

SNPs that are associated with disease are more likely to be eQTLs. A study by Nicolae *et al* (Nicolae *et al*. 2010) used a LCL model for autoimmune disease to compare eQTLs. The results showed an increased enrichment of eQTLs in the autoimmune disease group compared to other groups. When the eQTL data was used to annotate associated SNPs undiscovered loci were identified in complex disorders, helping to refine GWAS signals. Further evidence of a relationship between eQTL and GWAS SNPs was found in a recent study (Wright *et al*. 2014) which used the gene expression profiles of 2752 twins to quantify eQTLs in blood. Genotyping of 2494 twins identified eQTLs which were replicated in 1,895 unrelated twins and showed that the eQTLs overlapped with GWAS SNPs. Over 90% of the associated SNPs were in non-coding regions, with ~77% located in deoxyribonucleaseI (DNaseI) hypersensitive sites.

DNase Sensitivity eQTLs (ds-eQTLs) have been mapped using DNase-seq to identify regions of transcriptionally active, open chromatin (Degner *et al*. 2012). ds-eQTLs are variants which cause a reduction in chromatin accessibility, thereby affecting gene expression. ChIP-seq data from ENCODE, which looked at 9 transcription factors in one or more LCLs, was examined and it was found that alleles with increased DNaseI sensitivity had more transcription factor binding, so it was suggested that ds-eQTLs could be used as a predictor of transcription factor binding.

A recent study analysed 112,302 eQTL pairs using an eQTL browser and found that 80% of eQTLs were intra-chromosomal with the SNP at least 50kb from the boundary of associated genes (Duggal *et al*. 2014). eQTL data was mapped onto data that indicates the extent of chromatin interaction (Hi-C data) (Dixon *et al*. 2012) which found a significant relationship between eQTLs, target genes and chromatin structure. This study showed that eQTL fragments often interacted with other genomic fragments, were close to domain boundaries, close to target genes especially within domains, and were able to spatially associate with genes across domain. Further evidence that eQTLs can be placed close to distant genes has been shown in studies by Davison *et al*

(*DEXI*) (Davison *et al*. 2012), Meyer *et al* (*PVT1*) (Meyer *et al*. 2011) and Sotelo *et al* (*c-Myc*) (Sotelo *et al*. 2010).

To study eQTLs, either microarrays or RNA-seq can be employed (Majewski *et al.* 2011). RNA-seq has recently become more accessible and has confirmed previously generated microarray results (Montgomery *et al.* 2010; Pickrell *et al.* 2010). RNA-seq can be used to identify alternative splicing (sQTLs and isoform eQTLs) by mapping sequencing reads to splice junctions. The initiation and speed of transcriptional, mRNA processing, and post-transcriptional effects (mRNA stability) can all be studied by RNA-seq. Direct detection of cis-regulatory variation can be studied by allele counting whereby alleles closer to the gene have the strongest effect. Tissue specific effects have been identified through both microarrays and RNA-seq and have been used to create eQTL databases such as GeneVar (Yang *et al.* 2010), GEO (Edgar *et al*. 2002), Blood eQTL browser (Westra *et al*. 2013), and MuTHER (Multiple Tissue Human Expression Resource) (Grundberg *et al.* 2012). The GTEx database (GTEx Project. 2013) has eQTL information for more than 60 tissues types.

Studying eQTLs can provide evidence that a SNP is having an effect on a particular gene, however, there are limitations to take into account. The catalogues of eQTL data are often incomplete, many eQTLs (especially *trans*-eQTLs) do not replicate (Verdugo *et al.* 2010; Verlaan *et al.* 2009; Xia *et al.* 2012), there is variation between tissue types and cells activated by different stimuli, batch effects can occur, and cell heterogeneity can all affect the results.

1.3.5. Epigenetics in autoimmunity

It has been proposed that epigenetic changes which can affect the regulation of gene expression can be brought about through environmental factors such as smoking, resulting in the activation of the immune system (Karlson *et al.* 2010). The main focus of epigenetics studies in RA has been the role of DNA methylation and post-translational histone modifications (Klein *et al.* 2012; Klein *et al.* 2015).

1.3.5.1. DNA methylation

DNA methylation is catalysed by a family of enzymes called DNA methytransferases (DNMT) (Fuks *et al.* 2000). DNMTs catalyse the methylation of CpG dinucleotides (cytosine and guanine separated by only one phosphate) at the carbon-5 position forming 5-methylcytosine (5-mC), which are clustered in regions of the genome known as CpG islands. CpG islands are enriched in active gene promoters and are generally hypomethylated. Therefore, methylation at these sites can result in transcription initiation being blocked causing gene silencing for example, in X-inactivation (Arand *et al.* 2012; Viatte *et al.* 2013).

Changes in DNA methylation has been proposed as a factor adding to the genetic risk in RA (Liu *et al.* 2013). A number of epigenetic studies support this theory and have indicated that the methylation pattern in cells such as CD4+ T-cells and synovial fibroblasts from patients with RA and other autoimmune diseases is altered.

Global hypomethylation has been observed in T-cells and peripheral blood cells from RA patients (Karouzakis *et al.* 2009). Hypomethylation in CD4+ T-cells results in decreased expression of genes such as *DNMT1* (Lei *et al*. 2009; Richardson *et al*. 1990). Reducing the amount of methylation on genes important in immune cell stimulation leads to increased immune response (Liao *et al.* 2012) and increased activity of genes in pathways involved in cell migration (Nakano *et al.* 2013). Peripheral blood mononuclear cells (PBMCs) from RA patients were shown to be demethylated at a single CpG site in the *IL-6* promoter (encoding IL-6) which is a critical factor in B-cell response (Nile *et al.* 2008).

Hypermethylation has also been implicated in gene regulation in various cell types. In CD4+ T-cells (Wang *et al.* 2014) hypermethylation affected the expression of a key transcription factor responsible for the generation of regulatory T-cells, and in synovial fibroblasts genes important in apoptosis and in the TGF-beta signalling pathway were silenced (Park *et al.* 2013; Takami *et al.* 2006).

As well as global effects, there is evidence that the promoters of specific genes in different cell types can be differentially methylated. Examples of specific genes include *EBF3* (early B-cell factor 3) and *IRX1* (Iroquois homeobox 1) in synovial fibroblasts (Park *et al.* 2013), and the miR-124a gene promoter (Zhou *et al.* 2013a). It has also been shown that a whole cell population can be affected by a very small change in the methylation pattern, for example, regulatory T-cell (T_{REG}) function has been shown to be compromised by methylation of the *CTLA-4* promoter at just a single site (Cribbs *et al.* 2014).

1.3.5.2. Histone Modifications

Histone modifications are another example of epigenetic changes associated with RA. Histones are responsible for packaging the DNA into nucleosomes and post-translational modifications allow the DNA to be further allocated into regions of euchromatin, where the DNA is open and accessible for transcription, or heterochromatin where the DNA is tightly packed and not transcriptionally active. Histones contain two subunits each of core histones H2A, H2B, H3 and H4, and stabilising histones H1 and H5.

Histone methylation is catalysed by the enzymes histone methyltransferases (HMT) and histone demethylases (HDM) (Klein *et al.* 2015; Kouzarides 2002; Kouzarides 2007). Methylation can be mono-methylation, di-methylation or tri-methylation. Mono-methylation of the fourth lysine of histone H3 (H3K4me1) signifies an active enhancer and is enriched downstream of transcription start sites, di- or tri-methylation at the same site (H3K4me2 and H3K4me3) is associated with active or poised promoters (ENCODE Project 2012; Kouzarides 2007). Tri-methylation of lysine 27 on histone H3 (H3K27me3) is a repressive mark, indicative of promoters that are silenced by Polycomb proteins (Kouzarides 2007).

Acetylation is carried out through the action of histone deacetylases (HDAC) and histone acetyltransferases (HAT) (Klein *et al.* 2015; Kouzarides 2007). A recent study in RA synovial fibroblasts has shown that histone deacetylase 1 (HDAC-1) was more highly expressed in RA cells compared to synovial fibroblasts from osteoarthritis patients and upregulated genes involved in cell migration, proliferation and invasion, suggesting that HDAC-1 could play an important role in RA pathogenesis (Hawtree *et al.* 2015). Acetylation of lysine 27 on histone H3 (H3K27ac) differentiates between active and inactive promoters and enhancers. De-acetylation represses gene expression (Kouzarides 2007) and global hypoacetylation has been observed in the CD4+ T-cells (Hu *et al.* 2008a) of SLE patients.

The regulation of specific genes can also be altered through differential histone modification, for example, the *CD70* gene. CD4+ T-cells from patients with SLE have higher levels of H3K4me2 and H3 acetylation resulting in upregulation of *CD70* gene expression and an increased immune response (Zhou *et al.* 2011). Increased levels of histone methyltransferase have been shown to silence expression of a gene responsible for controlling collagen deposition by synovial fibroblasts (Trenkmann *et al.* 2011).

Further examples of epigenetic changes include chromatin remodelling and non-coding ribonucleic acids (ncRNAs) which are discussed below (Fulci *et al*. 2010).

1.3.6. Non-coding RNAs in immunity

Contrary to popular belief, approximately 75% of the human genome is transcribed (Djebali et al. 2012). However, only around 2% of these transcripts are translated into functional proteins (Zhang *et al.* 2015b) with long non-coding RNAs (lncRNAs) making up the vast majority of transcripts. lncRNAs are at least 200 nucleotides in size (Rinn *et al.* 2012), are polyadenylated and can be spliced into transcripts having +1 exon but are shorter than mRNA.

Over 10,000 lncRNAs have been identified, of which the majority are classified as long intergenic non-coding RNA (lincRNA). The most common class of lincRNA are the enhancer RNAs (eRNA) (Lam *et al.* 2014; Mousavi *et al.* 2014) which are correlated with expression of neighbouring protein-coding genes (Vance *et al.* 2014). Other classes of lncRNA include intronic lncRNA, antisense lncRNA, and transcribed pseudogene lncRNA (Zhang *et al.* 2015b). It has been shown that a single lncRNA can interact with multiple binding partners, over large genomic distances and even on different chromosomes. For example, HOTAIR lncRNA is transcribed from the HoxC locus and associates with hundreds of binding locations across many chromosomes. In particular, HOTAIR represses the transcription of the HoxD gene cluster on a different chromosome (Vance *et al.* 2014).

Many biological processes can be regulated by lncRNAs, for example, the differentiation and activation of immune cells and have been comprehensively reviewed (Fitzgerald *et al.* 2014; Zhang *et al.* 2015b). The diverse molecular functions include chromatin modification, transcriptional co-activation, regulation of translation, RNA turnover and splicing (Fitzgerald *et al.* 2014).

Another example of non-coding RNA are the microRNAs (miRNA), reviewed in (Bulik-Sullivan *et al*. 2013), which are approximately 22 nucleotides in length, transcribed by RNA Polymerase II, non-coding, and can regulate gene expression at the post-transcriptional level (Bartel 2009). miRNAs regulate gene expression through binding to complementary sequences on mRNA leading to translational repression, destabilisation and degradation of mRNA (Zhang *et al.* 2015b).

miRNAs were first associated with disease in a muscular hypertrophy in Texel sheep (Clop *et al*. 2006). In human disease, a synonymous variant has been identified that alters a miR-196 target site and influences the risk for Crohn's disease (Georges 2011). miR-21 expression has been linked to a number of autoimmune diseases including SLE where miR-21 expression is increased in CD4+ T-cells (Zhang *et al.* 2015b). miR-21 targets an autoimmunity gene, *RASGRP1* (RAS Guanyl Releasing Protein 1 [Calcium And DAG-Regulated]), which regulates the Ras-MAPK pathway resulting in downregulation of *DNMT1* and DNA hypomethylation (Pan *et al.* 2010). It has been recently reported that miR-573 is a negative regulator in RA pathogenesis (Wang *et al.* 2015). In this study the authors found that *TXNDC5* (thioredoxin domain containing 5), which had previously been shown to be upregulated in RA synovial tissue, was directly targeted by miR-573 along with TLR-2 and EGF receptor. miRBase (Griffiths-Jones et al. 2006) is a bioinformatics pipeline to catalogue and prioritise variants in the miRNA regulome as functional candidates.

1.3.7. Studying the functional effect of SNPs

In order to study the functional effect of SNPs bioinformatic analysis and laboratory-based investigations, either *in vitro* or *in vivo*, can be carried out. Bioinformatic analysis using publicly available data can be used to determine if a SNP is likely to change a protein function, or whether an intergenic region containing an associated variant shows evidence of regulatory activity (Table 1). Markers of regulatory elements such as modified histones and transcription factors have been mapped extensively in dozens of cell lines and also some primary cells using chromatin immunoprecipitation (ChIP) followed by high throughput sequencing (ChIP-seq) by the ENCODE project (ENCODE Project 2012). Polymorphisms that affect binding of regulatory proteins can have a profound influence in disease, as the differences in binding that lead to downstream differences in expression may be the underlying cause of the disease associated SNPs (Schaub *et al*. 2012). The effects of polymorphisms can be investigated in the laboratory using techniques that study DNA-DNA or DNA-protein interactions.

There is a requirement post-GWAS to define the disease associated variants that may change transcription regulatory elements, the mechanism of regulation and the genes they influence. Although in linear view it may appear that disease associated SNPs are located far away from genes, there is well established evidence that within cells, the 3-D conformational structure of DNA often means that distant genomic regions are brought into close proximity (Davison *et al.* 2012; Pomerantz *et al.* 2009; Zhang *et al.* 2012a). Evidence that DNA containing associated markers is in molecular contact with distal genes can give confidence that the correct causal gene has been

identified. Techniques such as chromosome conformation capture (3C) can provide such evidence.

Table 1: Database and laboratory techniques used to investigate the links between SNPs and functional effects

Putative SNP function	Database	Laboratory technique
Dysregulation of protein structure/function	Polyphen-2	Immunoblot
		Microscopy
		Reporter gene assay
		siRNA knockdown
		Biophysical techniques
Modulation of gene expression	GeneVar	eQTL
	GTEx	qRT-PCR
	BioGPS	RNA-seq
Disruption of DNA-protein interactions	TRANSFAC	EMSA
	JASPAR	ChIP
	ENCODE	ChIA-PET
Chromatin structure	ENCODE	ChIP-seq
	NIH roadmap epigenomics project	
3-D Chromatin interactions	ChIA-PET browser	3C technologies (3C, 4C, 5C, Hi-C)
	Umass 5C	
		Microscopy (FISH)
Non-coding RNA	miRBase	RNA-seq
Epigenetic changes	GeneVar	EWAS
	NIH roadmap epigenomics project	

1.4. Investigation of DNA-DNA interactions

The interaction of long-range enhancers with their target genes is likely to be key to understanding how genetic variants influence complex genetic diseases, and the investigation into these interactions has been fundamental to my PhD.

1.4.1. Genome organisation

There is increasing evidence that interactions between regulatory regions of the genome play an important role in the regulation of gene expression, therefore studying how the mammalian genome is packaged in the nucleus is crucial to understanding the regulation of gene expression, reviewed in (Gibcus *et al.* 2013; Lanctot *et al.* 2007), and summarised in Figure 5.

Chromatin is a complex consisting of DNA, histones and accessory proteins. At the first level of organisation, chromatin is packaged into nucleosomes each containing 146bp of DNA wrapped 1.65 times around a histone octamer containing two copies each of histones H2A, H2B, H3 and H4 (Rodriguez *et al.* 2013). The DNA-nucleosome complex is arranged as a 10nm fibre which under certain conditions can form higher-order 30nm helical fibres (Hubner *et al*. 2013).

Within nucleosomes, chromatin organisation is guided through contact with the nuclear envelope and nuclear lamina. Within the nucleus, the chromatin is packaged into spatially separate chromosome territories containing loosely packed, active euchromatin or condensed, inactive heterochromatin (Rodriguez *et al.* 2013) referred to as Compartment A and Compartment B respectively (Fraser *et al.* 2007; Lieberman-Aiden *et al.* 2009; Zhang *et al.* 2012d) which are surrounded by inner and outer nuclear membranes (Cremer *et al.* 2010). Individual chromosomes are located within defined chromatin territories, organised with gene-rich regions orientated towards the nuclear interior and regions with fewer genes orientated towards the outer areas (Cremer *et al.* 2006a; Cremer *et al.* 2010). Repositioning of genomic regions, which takes place during a very small time window during the G1 interphase of mitosis, is thought to be important for the regulation of gene expression (Lanctot *et al.* 2007; Naumova *et al.* 2013).

Within chromatin territories, looping interactions take place between proximal promoters and distal regulatory elements such as enhancers, silencers and insulators (Dixon *et al.* 2012; Gibcus *et al.* 2013; Levine *et al.* 2014). A single promoter can interact with several enhancers leading to differential gene expression (Bulger *et al.* 2011). Insulator elements play a role in long-range chromatin looping and are bound by CCCTC-binding factor (CTCF) at CTCF binding sites (CBS) which require the recruitment of the cohesin complex for activity (Ong *et al.* 2014).

The genome can be further organised into cell-specific, topologically associating domains (TADs) (Dixon *et al.* 2012; Phillips-Cremins *et al.* 2013) which contain sequences that preferentially interact with each other rather than with other regions of the genome. TADs are spatially separated by boundary regions enriched in CTCF sites and housekeeping genes which act as insulators, blocking interactions between adjacent TADs (Dixon *et al.* 2012). Evidence that chromatin

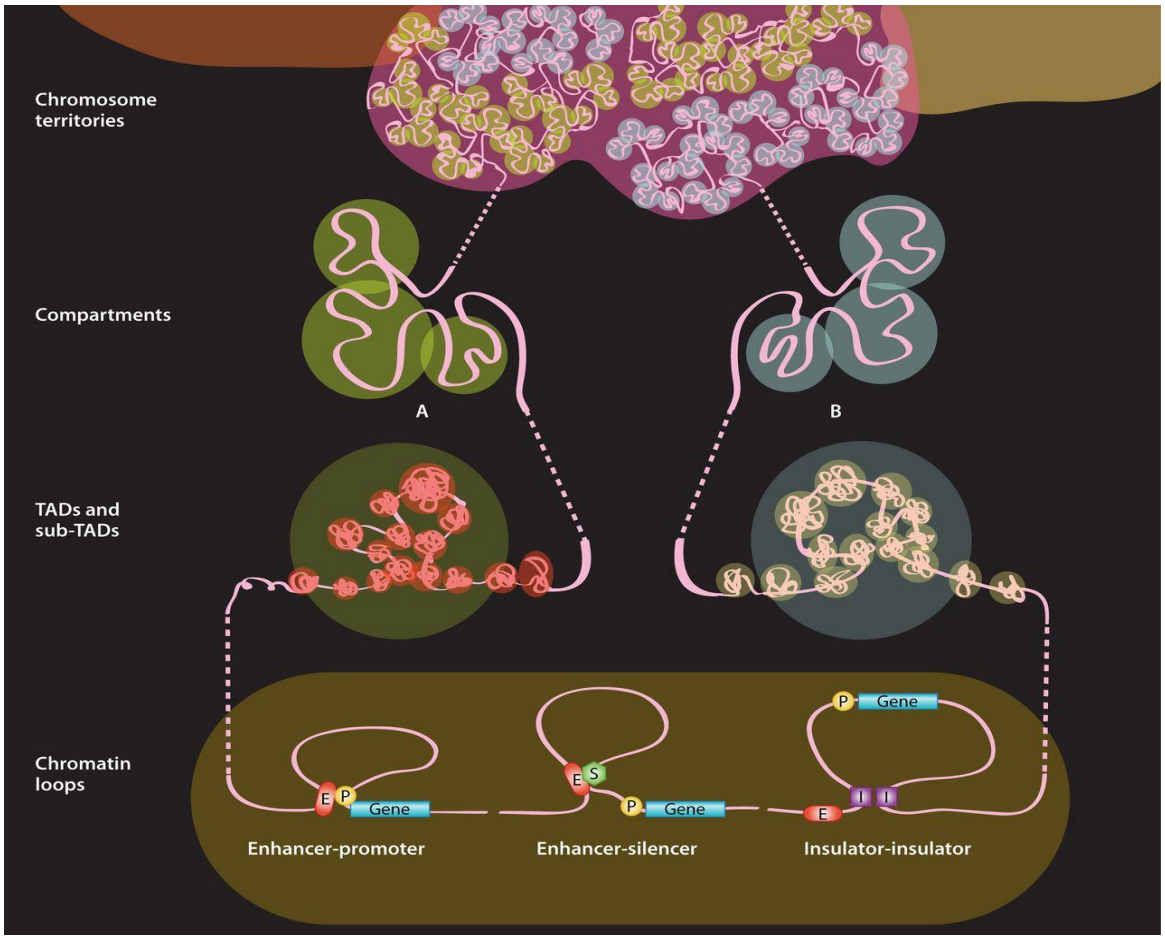
organisation is linked to gene expression comes from a number of studies. It has been reported that promoter-enhancer interactions occur preferentially within TADs (Jin *et al*. 2013), and that enhancer activity is strongly associated with TADs suggesting that TADs organise regulatory activities into large domains contributing to specific gene expression profiles (Symmons *et al.* 2014). A recent study (Guo *et al.* 2015) showed that CTCF binding sites are arranged in a forward-reverse orientation and if the sites are inverted using CRISPR/Cas9 gene editing, genome folding is altered changing enhancer/promoter function. Lupianez *et al* (Lupianez *et al.* 2015) found that disruption of TADs can contribute to pathogenic phenotypes. CRISPR/Cas9 genome editing was used to generate mice containing mutations linked to limb malformations showing that deletions, inversions and duplications in a TAD spanning a number of critical genes were responsible for the phenotype.

Genomic loci that are far apart can associate to form long-range interactions that can be intra- or inter-chromosomal (Miele *et al.* 2006; van Steensel *et al.* 2010). Loci within the same chromosome territory, and chromosomes with similar size and density, more frequently interact than loci in different chromosome territories (Kalhor *et al.* 2012; Lieberman-Aiden *et al.* 2009).

Long-range interactions have been shown in many studies to be involved in transcriptional regulation (Fraser *et al.* 2007; Smallwood *et al.* 2013). For example, a study of the Polycomb group of proteins (PcG) in *Drosophila melanogaster* showed that PcG proteins bind to *cis*-acting elements some distance away allowing the PcG proteins to act as repressors (Bantignies *et al.* 2006; Sexton *et al.* 2012b). Transcriptional activation in the β-globin locus has been shown to be regulated by the binding of locus control regions (LCRs) to downstream regulatory elements by the formation of a 200kb loop in cells expressing the genes (Tolhuis et al. 2002). Recent studies have shown that genes can associate with 'transcription factories' enriched in RNA polymerase II or KLF1 (kruppel-like factor 1 [erythroid]), for example, in mouse erythroid progenitor cells, where several genes can associate with the same enriched area (Osborne *et al.* 2004; Schoenfelder *et al.* 2010a; Schoenfelder *et al.* 2010b).

Most of the discoveries relating to genome organisation over the last decade have been due to the development of techniques such as Chromosome Conformation Capture (3C) which allow a high-resolution view of chromatin interactions (de Wit *et al.* 2012).

Figure 5: Chromatin organisation, showing the different levels of packaging within the nucleus (Illustration from James Fraser et al. Microbiol. Mol. Biol. Rev. 2015;79:347-372)



1.4.2. The development of 3C technologies

3C technologies are based on the principles of proximity ligation, whereby loci in close physical proximity are more likely to interact. All of the variations of 3C share some of the experimental steps, summarised in Figure 6 (Duan *et al.* 2012; Simonis *et al.* 2007). Depending on the application, different methods to detect interacting fragments can be used, such as PCR (polymerase chain reaction), qPCR (quantitative PCR), microarray or next generation sequencing (NGS). The different technologies are comprehensively reviewed in (de Laat *et al.* 2012; Dostie *et al.* 2012; Duan *et al.* 2012; Ethier *et al.* 2012; Fullwood *et al.* 2009b; Sexton *et al.* 2009) and summarised in Table 2.

Whilst 3C and related technologies are powerful techniques and crucial to enhancing the knowledge of chromatin organisation, there are limitations (Duan *et al.* 2012; Ethier *et al.* 2012; Lanctot *et al.* 2007; Simonis *et al.* 2007). Due to the low frequency of long-range interactions the number of cells needed to detect interactions is very high (Ethier *et al.* 2012), a minimum of $1x10^7$ cells, which also makes for limited throughput and low signal-to-noise ratio within experiments. Interactions between neighbouring fragments are more likely to occur, therefore, large volume

reactions are required to counter the over-representation of local interactions, leading to complicated, time-consuming and expensive protocols (Duan *et al.* 2012). The need for robust controls to eliminate false-positives and the prior knowledge of the interacting regions needed to enable the design of sequence-specific primers also adds to the complexity and limitations of the experimental procedures (Ethier *et al.* 2012). To overcome the limitations of the 3C process, modifications of the protocol are being continuously developed to make the protocol less complicated to perform, improve signal-to-noise ratios and improve sensitivity, summarised in Figure 6.

1.4.2.1. Chromosome conformation capture (3C)

The 3C protocol is commonly used to confirm specific physical interactions between a pair of loci and determine the relative frequency of the interactions (Dekker *et al.* 2002; Naumova *et al.* 2012; van Steensel *et al.* 2010). The following steps are used to produce a library of interacting fragments throughout the genome. Firstly, DNA-protein crosslinks are fixed using formaldehyde (FA) which forms covalent crosslinks between the primary amino groups of lysine and arginine side chains (Ethier *et al.* 2012; Jackson 1999). Following fixation of the cells, restriction enzymes are used to digest the DNA into regularly-sized fragments. Restriction enzymes that recognise a 6-bp sequence, such as *Hin*dIII, are the most commonly used but 4-bp cutters can also be used to increase resolution (Simonis *et al.* 2007). Ligation of interacting fragments is carried out under dilute conditions which are favourable towards intra-molecular ligation and then the crosslinks are reversed and DNA purified. Ligation products can then be detected one at a time using PCR or qPCR for specific genomic regions (Hagege *et al.* 2007; Naumova *et al.* 2012).

The 3C technique was developed in 2002 by Job Dekker (Dekker *et al.* 2002) to study the 3-D conformation of the *Saccharomyces cerevisiae* genome and has since been modified to study mammalian genomes (Miele *et al.* 2006). The first study to identify a long-range physical interaction between a gene and regulatory element using 3C was carried out in 2002 (Tolhuis *et al.* 2002), which showed that β-globin genes physically interacted with the locus control region (LCR), an interaction that was later found to be mediated by CTCF (Splinter *et al*. 2006). Spilianakis *et al* (Spilianakis *et al*. 2005) demonstrated long-range interactions between LCRs and the promoters of cytokines in T-helper (T_H) cells. Using 3C, it was found that IFN-γ on chromosome 10 physically interacted with the T_H2 cytokine locus on chromosome 11.

Table 2: Summary of the chromosome conformation capture technologies

Technology	Interaction	Detection	References
3C Chromosome conformation capture	One-to-one	PCR/qPCR	(Dekker *et al.* 2002; Naumova *et al.* 2012)
4C Chromosome conformation capture on chip (circular 3C)	One-to-many	Microarray/NGS	(Sexton *et al.* 2012a; Simonis *et al.* 2006)
5C Chromosome conformation capture carbon copy	Many-to-many	Microarray/NGS	(Dostie *et al.* 2006; Dostie *et al.* 2007a)
Hi-C	Genome wide	NGS	(Belton *et al.* 2012; Lieberman-Aiden *et al.* 2009; van Berkum *et al.* 2010)
TCC Tethered conformation capture	Genome wide	NGS	(Kalhor *et al.* 2012)
3C-seq (multiplexed 3C-seq)	Many-to-all	NGS	(Stadhouders *et al.* 2013)
ChIA-PET Chromatin interaction analysis with paired-end tag sequencing	Genome wide	NGS	(Heidari *et al.* 2014)
DNase Hi-C	Genome wide	NGS	(Ma *et al.* 2015)
T2C Targeted Chromatin Capture	Many-to-all	NGS	(Kolovos *et al.* 2014)
Capture-C	Many-to-many	NGS	(Hughes *et al.* 2014)
Capture Hi-C Use of RNA baits to target specific genomic regions	Genome wide	NGS	(Dryden *et al.* 2014; Jager *et al.* 2015; Martin *et al.* 2015; Mifsud *et al.* 2015)
HiCap	Genome wide	NGS	(Sahlen *et al.* 2015)

PCR, polymerase chain reaction; qPCR, quantitative PCR; NGS, next-generation sequencing

The development of 3C has provided valuable insight into interactions with risk loci identified in GWAS studies. The 8q24 region has been associated with cancers (Ahmadiyeh *et al*. 2010) such as colorectal (Tomlinson 2012), prostate (Haiman *et al*. 2007) and breast cancer (Bertucci *et al.* 2012; Li *et al.* 2011). An example of a risk locus in this region is *cMyc* which is a regulator of cellular growth, proliferation and apoptosis (Yochum 2011). Pomerantz *et al* (Pomerantz *et al.* 2009) used 3C to demonstrate that an enhancer region in the 8q24 region, which is associated with colorectal cancer, physically interacted with the *Myc* locus ~335kb away from the risk region. Investigating the same locus, Wright *et al* (Wright *et al*. 2010) found that a cancer-associated, intergenic SNP (rs6983267) in the 8q24 region interacts with the *cMyc* promoter by forming a 335kb loop. Other long range interactions with *Myc* promoters in this region were found to be tissue specific (Ahmadiyeh *et al.* 2010). Yochum *et al* (Yochum 2011) used 3C to identify that 5 novel enhancers in the 8q24 region, 400kb upstream from the *cMyc* TSS, formed long range loops that positioned them next to the *cMyc* promoter. The 16q12.1 locus associated with breast cancer that relapses to the bone has also been studied (Cowper-Sal.lari R. *et al*. 2012). It was found by 3C that the region containing the rs4784227 SNP physically interacted with the promoter region of the *TOX3* gene (TOX High Mobility Group Box Family Member 3), which modifies chromatin structure. This interaction resulted in allele-specific differences in chromatin affinity for *FOXA1* (Forkhead Box A1) at regulatory sites.

In autoimmune disease research, 3C has been used to interrogate the 16p13 region associated with several autoimmune diseases including T1D and multiple sclerosis (MS) (Davison *et al.* 2012). In this study, a novel long-range interaction was identified that formed a 15-kb loop between the promoter region of the *DEXI* gene (Dexamethasone Induced) and intron 19 of *CLEC16A* (C-Type Lectin Domain Family 16, Member A), expressed in immune system cells. In SLE, a chromatin loop was identified on chromosome 6q that enabled a physical interaction between variants residing in an enhancer element that binds to NF-κB and the *TNFAIP3* promoter which is an important modulator of immune activity (Wang *et al*. 2013).

It is clear that 3C has been successful in identifying novel long-range interactions in many studies but the main limitations of 3C are that prior knowledge of potential interacting regions is required in order to design specific primers, many controls are needed, and it is difficult to identify genuine interactions because the signal-to-noise ratio is very low (Dekker 2006; Kolovos *et al.* 2014). To address some of the limitations and improve specificity, modifications of 3C that use streptavidin-biotin pulldown and NGS have been developed (Duan *et al.* 2012), such as tethered conformation capture (TCC) (Kalhor *et al.* 2012), 3C-seq (Stadhouders *et al.* 2013) and very recently, high-throughput Capture C (Hughes *et al.* 2014).

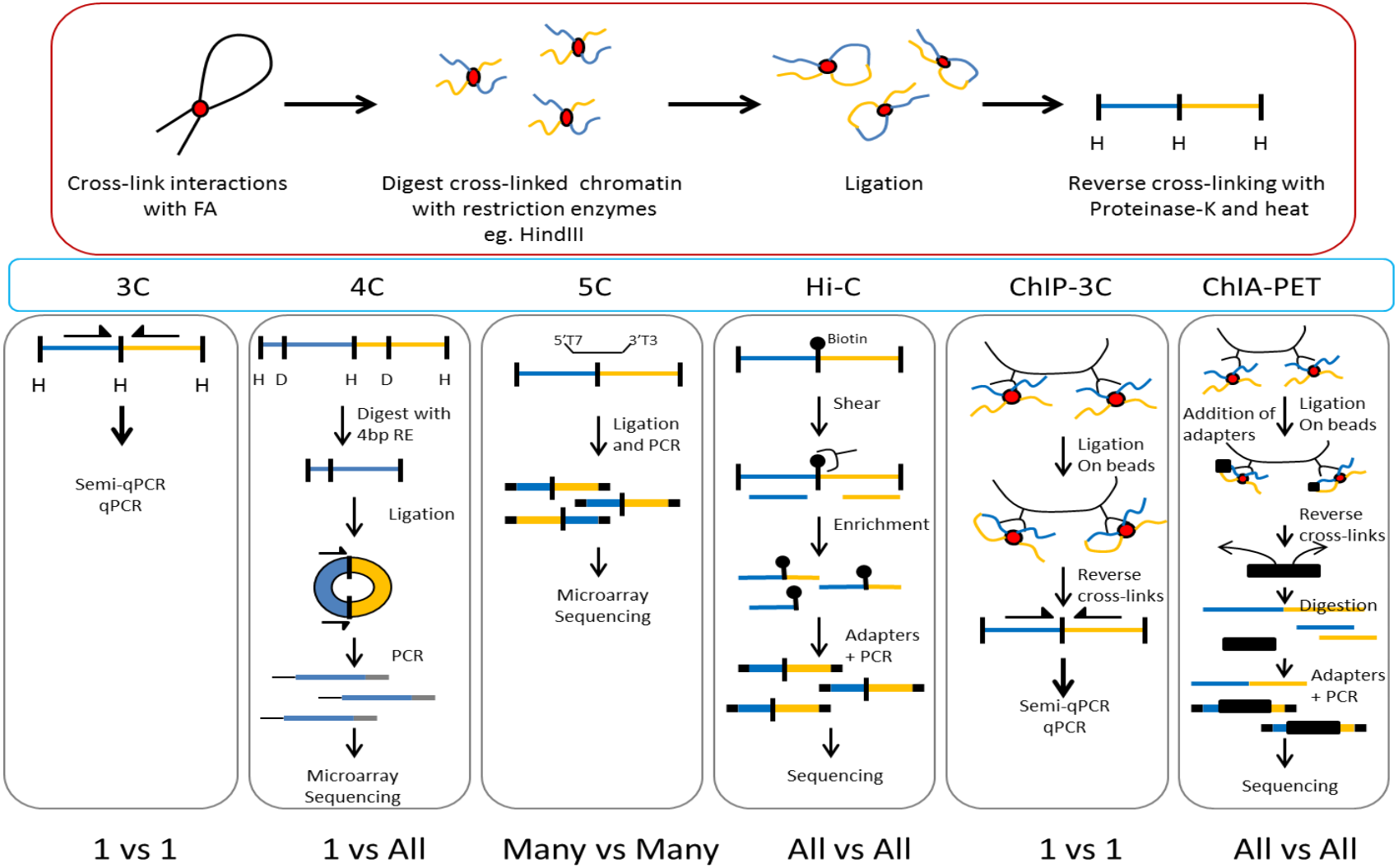
1.4.2.2. Chromosome conformation capture on chip (4C)

Simonis *et al* developed the 4C technique in 2006 to study the β-globin locus and *Rad23a* interaction profiles in erythroid cells (Simonis *et al*. 2006). The technique is also known as circular chromosome conformation capture (Zhao *et al*. 2006). Briefly, a 3C library is prepared using a 6-cutter restriction enzyme such as *Hin*dIII then digested using a second, frequently cutting enzyme such as *Dpn*II. Ligation is carried out under dilute conditions to circularise the fragments, then the circles are re-linearised to enable PCR amplification using a third restriction enzyme that recognises a site between the first and second restriction enzyme. The 4C libraries are amplified by PCR using target-specific primers then labelled and hybridised to microarray slides containing probes representing different restriction fragment ends throughout the region of interest. This variation of 3C can enable the study of all the regions interacting with a genomic site of interest (one versus all) (Fullwood *et al.* 2009b). Modified 4C protocols encompassing a ChIP step to pull down specific interacting fragments have also been developed (Apostolou *et al.* 2008; Schoenfelder *et al.* 2010a; Sexton *et al.* 2012a).

4C has been used to characterise the mouse hypersensitive site (HS2) in the β-globin LCR on chromosome 7 (Simonis *et al.* 2006). In this study it was found that the majority of the interactions were within the same chromosome territory, in a region centred around β-globin. Further studies on the β-globin locus have identified more interactions and confirmed previously identified interactions, such as with CTCF (Splinter *et al.* 2006; van de Werken *et al.* 2012b). The α-globin gene cluster has also been investigated using 4C, which found that when the α-globin gene cluster is expressed, physical interactions cause the expression of genes in a 500kb region around the cluster to increase (Lower *et al*. 2009).

Other groups have used 4C followed by detection of interactions by NGS to increase resolution (Apostolou *et al.* 2013; Dermitzakis *et al.* 2005; Splinter *et al.* 2012; van de Werken *et al.* 2012a). Non-coding sequences make up the majority of the human genome and conserved non-coding (CNC) sequences are thought to have a functional role as they have been maintained through evolution (Dermitzakis *et al.* 2005). Robyr *et al* (Robyr *et al*. 2011) sequenced 4C libraries for 10 CNC regions in duplicate in the K562 myelogenous leukaemia cell line and found that 9 intergenic CNCs, and other interactions, were located within a 700kb region in chromosome 21 that contained the *OLIG1* and *OLIG2* genes.

Figure 6: Comparison of the different 3C technologies



1.4.2.3. Chromosome conformation capture carbon copy (5C)

The 5C technique is used to study long-range interactions using a 'many vs many' strategy, meaning that several loci can be interrogated simultaneously. The 5C protocol was developed by Dostie *et al* (Dostie *et al.* 2006; Dostie *et al.* 2007a; Dostie *et al.* 2007b) to further characterise the human β-globin locus, with a recent update to the methodology in 2012 (Ferraiuolo *et al*. 2012). 3C libraries are prepared and then ligation mediated amplification (LMA) is used to copy and amplify parts of the 3C library, making a 'carbon copy'. LMA works by detecting target sequences that are amplified using primers that anneal next to each other on the DNA strand. 5C primers that are annealed next to each other are ligated using Taq ligase and the library amplified using universal primers that anneal to the ends of the 5C primers. The interacting fragments can be detected using NGS or microarray.

Sanyal *et al* (Sanyal *et al*. 2012) used both 5C and 3C to generate a long-range interaction landscape of gene promoters. 5C was used to map interactions between promoters and distal elements throughout 44 ENCODE (ENCODE Project 2012) regions representing 1% of the human genome (30Mb) in 3 cell lines - GM12878, K562 and HeLa-53. Interactions were followed up with a targeted 3C approach using 981 reverse primers targeting the TSS and 5,321 forward primers targeting distal regulatory regions. Over 1000 long-range interactions were detected in each cell line, with interactions between TSS and ~120kb upstream regions being the most common. Interestingly, only a small percentage of interactions were with the nearest gene (~7%).

Phillips-Cremins *et al* (Phillips-Cremins *et al*. 2013) found that the 3-D organisation of the mammalian genome during lineage differentiation of stem cells was shaped by different combinations of proteins. 5C in conjunction with NGS generated a high resolution (~4kb) map of interactions, identifying ~90,000 *cis* and ~500,000 *trans* interactions which showed that the proteins CTCF, Mediator and Cohesin were the main drivers of chromatin interactions during differentiation.

1.4.2.4. Hi-C

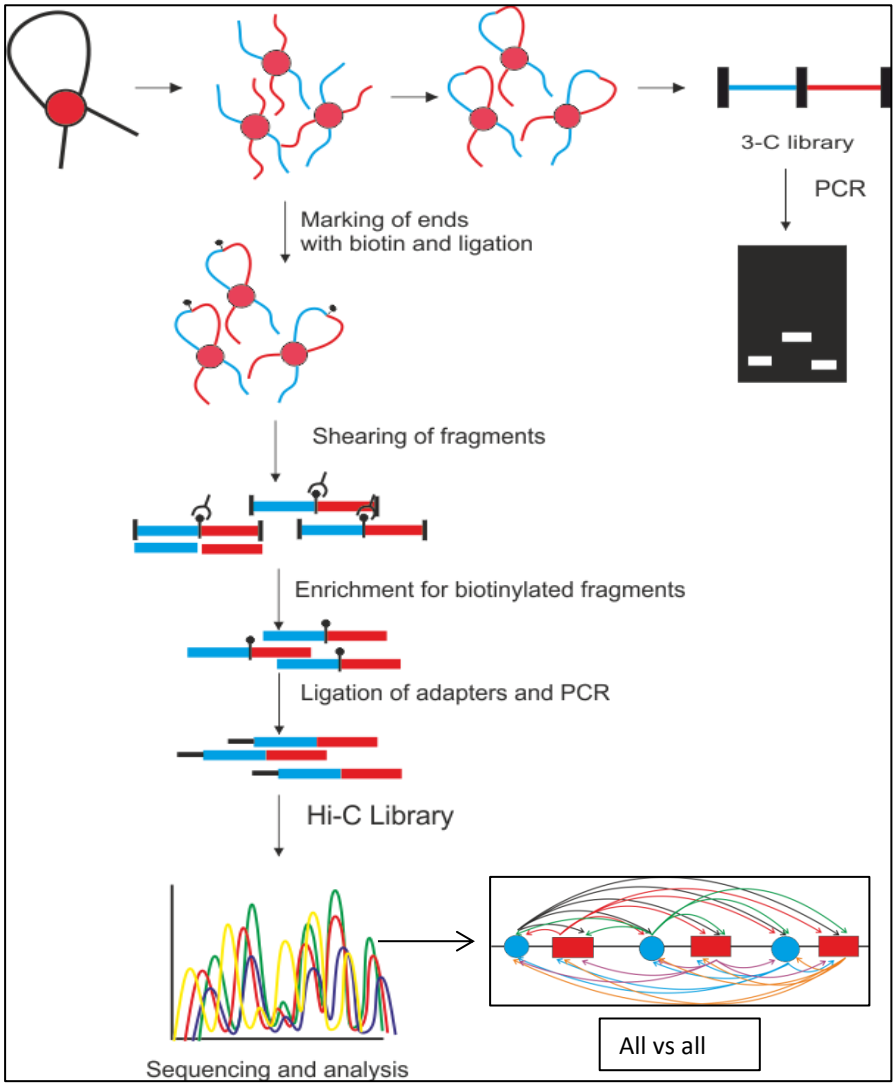
The Hi-C protocol is the most recent of the 3C technologies. The protocol was developed in 2009 by Lieberman-Aiden *et al* (Lieberman-Aiden *et al.* 2009) and enables the genome-wide study of higher-order chromatin interactions by coupling 3C with high-throughput NGS (Hi-C) (Figure 7). To make a Hi-C library, the chromatin is digested as for a 3C library then the digested ends are marked with biotin prior to ligation allowing the selective enrichment of ligation junctions using streptavidin beads. Adapters are ligated to the Hi-C fragments to enable the detection of interactions by paired-end NGS. The Hi-C methodology has been described in detail by van Berkum *et al* (van Berkum *et al*. 2010) and Belton *et al* (Belton *et al*. 2012).

In the original study (Lieberman-Aiden *et al.* 2009), Hi-C was carried out on a human lymphoblastoid cell line (LCL), GM06990, and the K562 erythroleukaemia cell line. An Illumina Genome analyser was used to generate 8.4 million read-pairs, of which 6.7 million corresponded to long-range interactions (LRI) between regions >20kb apart. The genome was split into 1Mb

regions and a matrix of interactions generated and visualised on a heat-map indicating interaction frequency. The results from Hi-C were consistent with findings previously generated in 3C and FISH (fluorescent *in situ* hybridisation) studies, showing that chromosomes within the same chromosome territory interact more frequently than with chromosomes in different territories (Cremer *et al.* 2006b; Ethier *et al.* 2012; Gibcus *et al.* 2013; Kalhor *et al.* 2012; Lieberman-Aiden *et al.* 2009).

The 3-dimensional organisations of the *D.melanogaster,* mouse and human genomes have all been studied at high resolution using Hi-C (Belton *et al.* 2012; Dixon *et al.* 2012; Lieberman-Aiden *et al.* 2009; Sexton *et al.* 2012b). In mouse and human genomes, large areas of interacting chromatin known as topological domains were identified, which are enriched in CTCF and housekeeping genes (Dixon *et al.* 2012).

Figure 7: The standard Hi-C protocol



In 2013, Jin *et al* (Jin *et al.* 2013) used Hi-C in human fibroblast cells to characterise the dynamics of promoter-enhancer interactions following TNF-α signalling. They found that TNF-α responsive enhancers were already in contact with their promoters before signalling took place, suggesting that long-range interactions were a strong predictor of gene induction. Also, the majority of the LRIs detected were located within the same chromosome territory and showed a preference for interactions with promoters over enhancers, which supports evidence that LRIs are important for transcriptional regulation (Smallwood *et al.* 2013).

Hi-C has been modified to enable the characterisation of long-range interactions in the individual nuclei of mouse T_H1 cells (Nagano *et al*. 2013). T_H1 cells were differentiated *in vitro* from $CD4^+$ T-cells isolated from mouse spleens then a modified Hi-C protocol was used to crosslink, digest, biotin-mark and ligate the chromatin inside the nucleus (the original Hi-C protocol carries out these steps following cell lysis). Individual nuclei were isolated under a microscope then crosslink reversal and isolation of biotinylated Hi-C junctions carried out in individual tubes. A second restriction digestion was performed and the fragments ligated to adapters. PCR was used to amplify the single-cell libraries which were characterised by NGS. Single-cell Hi-C revealed that chromosomes retain domain organisation on a small scale, but there is cell-to-cell variability on a larger scale.

Recent modifications to the Hi-C protocol have sought to increase the efficiency and resolution of the technique. In 2014, Rao *et al* (Rao *et al.* 2014) developed *in situ* Hi-C to generate a high resolution map of the human genome. This variation of Hi-C uses the original protocol but with the ligation step carried out in intact nuclei, which is similar to the Nagano *et al* single cell Hi-C protocol discussed above. Using *in situ* Hi-C generated a map of the human genome in nine human cell lines at 1kb resolution, providing a wealth of information about chromatin looping and genome organisation. The improved protocol also produced fewer random ligation products, was quicker and offered higher resolution through the use of a 4-cutter restriction enzyme. Nagano *et al* recently carried out a comparison of in-solution Hi-C to in-nucleus Hi-C (Nagano *et al.* 2015) and showed that the in-nucleus protocol gave less experimental noise, was simpler to perform and produced better quality libraries.

1.4.2.5. Capture Hi-C

Protocols that utilise a sequence capture step allow the interrogation of interactions between specific regions and the whole genome in an unbiased way. The first Hi-C protocols to use a capture step used capture arrays from Nimblegen (Roche) which is a tiling microarray which probes for specific DNA sequences in a particular region of the genome and has successfully been used in DNase-Chip and transcriptome mapping (Bertone *et al.* 2005; Scacheri *et al.* 2006). Targeted chromatin capture (T2C) (Kolovos *et al.* 2014) and HiCap (Sahlen *et al.* 2015) are both variations of Hi-C which use a sequence capture step to enrich for particular genomic regions. T2C was used to interrogate the mouse and human genomes at single restriction fragment resolution. HiCap used a 4-cutter restriction enzyme coupled with sequence capture of promoter regions

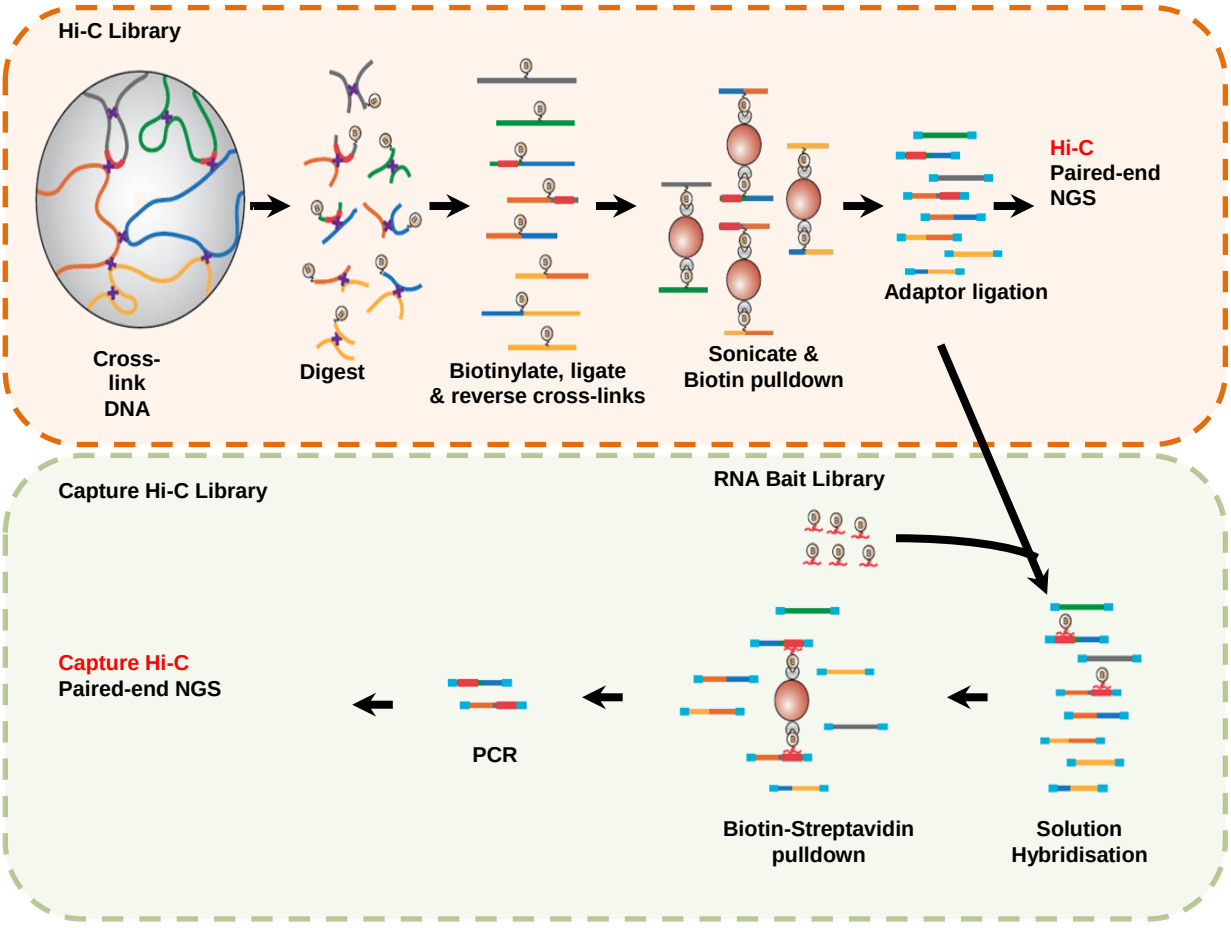
resulting in fragment sizes of 699bp, which is close to single enhancer resolution (Sahlen *et al.* 2015). DNase Hi-C (Ma *et al.* 2015) employed DNaseI instead of a restriction endonuclease to fragment the chromatin and was used to map the 3-D organisation of lincRNAs using Nimblegen and Illumina sequencing.

The recent development of Capture Hi-C (CHi-C) (Figure 8) uses solution hybridisation (Gnirke *et al.* 2009) followed by Illumina sequencing instead of capture arrays. Hi-C libraries are hybridised to custom-designed RNA baits (Dryden *et al.* 2014) allowing an unbiased genome-wide view of interactions with targeted genomic regions. This approach means that the sequencing depth at targeted regions is significantly increased over traditional Hi-C enabling the identification of specific interactions. CHi-C has been used recently to study breast cancer susceptibility loci (Dryden *et al.* 2014), colorectal cancer risk loci (Jager *et al.* 2015), and long-range interactions between promoters and their regulatory elements (Mifsud *et al.* 2015; Schoenfelder *et al.* 2015).

In autoimmune disease, a Capture Hi-C approach using complementary region and promoter capture experiments in B-cell and T-cell lines has recently been used by my group to identify novel target genes and complex long range interactions with related autoimmune risk loci (Martin *et al.* 2015).

Figure 8: The Capture Hi-C protocol

Adapted from Schoenfelder *et al*. 2015



1.5. Investigation of DNA-Protein interactions

DNA-protein interactions are involved in processes such as gene activation, chromosome organisation and DNA repair, therefore the ability to study the molecules involved is very important. Several techniques, both *in vitro* and *in vivo*, are available and are comprehensively reviewed in (Christova 2013) and (Dey et al. 2012), and summarised in Table 3. For example, a simple ChIP assay can be routinely used to identify specific transcription factor binding or histone modifications at specific genomic sites techniques. More complicated, sequencing-based techniques such as ATAC-seq and ChIA-PET can be used to map regions of transcriptionally active open chromatin, histone modifications and transcription factor binding sites on a genome-wide scale.

Table 3: Summary of the different ways DNA-Protein interactions can be investigated

Technology	Scope	Advantages/disadvantages
ChIP (Chromatin Immunoprecipitation)	Identify specific transcription factor binding to a genomic region of interest	Simple to perform Many variations to suit application Low signal:noise ratio Reliant on antibody specificity
ChIA-PET	Sequences bound by transcription factors detected by NGS	Reliant on antibody specificity
EMSA (Electrophoretic mobility shift assay)	Identify proteins bound to a specific genomic region	Use of radioisotopes for maximum sensitivity Fluorescent versions available
Supershift-EMSA	EMSA followed by Western blotting using a specific antibody to transcription factor of interest	Reliant on antibody specificity Confirmation by ChIP required
Proteome wide association studies (PWAS)	Use of mass-spectrometry to detect interactions	Assay many SNPs and identify transcription factors in one assay Confirmation by ChIP required
ATAC-seq (Assay for Transposase-Accessible Chromatin)	Sequences bound by transcription factors detected by NGS	Transposase incorporation, low cell numbers

1.5.1. ChIP-based technologies

Chromatin immunoprecipitation (ChIP) is a popular method for identifying protein factor binding to DNA sequences, involving the immunoprecipitation of protein-DNA complexes using specific antibodies (Collas *et al.* 2008; Sikes *et al.* 2009). There are many variations of the ChIP protocol available such as µ-ChIP (Dahl *et al.* 2008) and nano-ChIP (Adli *et al.* 2011) which allow ChIP from very low cell numbers, and ChIP-seq (Mercier *et al*. 2011) which allows the genome-wide analysis of DNA-protein interactions. The different types of ChIP assay are summarised in Table 4. Commercial ChIP kits are available from several companies, for example, the MagnaChIP© kit from Millipore.

The most common ChIP experiment is crosslinking ChIP (X-ChIP), which involves crosslinking of DNA-protein interactions with formaldehyde. Interacting fragments are immunoprecipitated using specific antibodies immobilised to magnetic beads and detected by PCR, qPCR, sequencing or microarray. The main disadvantages of ChIP are that the sensitivity of the assay is dependent on the specificity of the antibody and the amount of signal compared to non-specific binding is very low making it difficult to determine genuine interactions (Fullwood *et al.* 2009b). The advantage is that the technique is relatively straightforward to perform, and the availability of commercially developed kits can also help to yield consistent results. Development of assays such as ChIP-chip and ChIP-seq (Wei *et al.* 2006; Wu *et al.* 2006) allow the detection of genome-wide interactions.

To improve the specificity of ChIP, ChIA-PET was developed in 2009 to investigate long-range interactions involving ER-α (oestrogen receptor) (Fullwood *et al.* 2009a). This assay uses ChIP to isolate specific interacting regions then a linker sequence is introduced in the junction between the two fragments during proximity ligation. Ligation products can then be detected by paired-end sequencing (Fullwood *et al.* 2010; Zhang *et al.* 2012b).

Heidari *et al* (Heidari *et al.* 2014) generated a genome-wide map of interactions between regulatory elements in human cell lines (K562 and GM12878) using CHIA-PET targeting six different factors. Over 80% of the bound sites, including transcription start sites and enhancers, corresponded to DNaseI HS sites. The main components at bound sites, contributing to the 3-D chromatin structure, were shown to be cohesin, CTCF and ZNF143. Interactions between enhancers and promoters were shown to be cell-type specific. Distal and proximal regulatory networks showed different biological functions and structure, with proximal events enriched at housekeeping genes and distal events involved in dynamic events such as response to stimuli. Interactions with transcription start sites, transcribed regions or enhancers were correlated with high gene expression whereas interactions with CTCF and weak enhancers, correlated with moderate gene expression.

Table 4: Summary of the different ChIP-based assay

ChIP Assay	Interaction	Detection	References
X-ChIP (formaldehyde crosslinked)	Single protein with target genomic region	PCR, qPCR	(Christova 2013; Collas 2010; Heintzman *et al.* 2007)
N-ChIP (native ChIP)	Histone modifications and proteins tightly bound to target genomic region	PCR, qPCR	(O'Neill *et al.* 2003)
Re-ChIP (sequential ChIP)	Two or more proteins bound in close proximity to the target genomic region	PCR, qPCR	(Metivier et al. 2003)
Exo-ChIP (Exonuclease digestion after ChIP)	Histone modifications Single protein bound to target genomic region (high resolution mapping)	PCR NGS	(Rhee *et al.* 2012)
ChIP-chip (ChIP followed by microarray)	Histone modifications Single protein bound to target genomic region	Microarray	(Wu *et al.* 2006)
ChIP-seq (ChIP followed by NGS)	Histone modifications Single protein bound to target genomic region	NGS	(Mercier *et al.* 2011; Wei *et al.* 2006)
ChIA-PET	Single protein interacting with DNA	NGS	(Fullwood *et al.* 2009b; Fullwood *et al.* 2010; Zhang *et al.* 2012b)

1.5.2. *In vitro* assays for investigating DNA-Protein interactions

In vitro assays such as electrophoretic mobility shift assays (EMSA) are useful if the specific transcription factor is not known (Hellman *et al.* 2007) and are relatively simple to carry out. The change in migration during polyacrylamide gel electrophoresis (PAGE) of a complex compared to naked DNA can detect if a protein/protein complex binds to a DNA sequence of interest. EMSA followed by Western blotting can confirm the identity of a transcription factor and super-shift EMSA using specific antibodies can be used to detect allele-specific binding once the transcription factor has been identified. Proteome wide association studies (PWAS) use a mass spectrometry based approach (Butter *et al.* 2012) to assay multiple SNPs and identify transcription factors in one experiment. Results from *in vitro* assays such as EMSA and PWAS need to be validated using ChIP because of the risk of false positives (Edwards *et al.* 2013). The impact of a given SNP can be further studied by using reporter gene assays, whereby the region of interest is cloned into a promoter-driven reporter construct.

1.6. Aims of the project

GWAS have identified many loci that are involved in the pathogenesis of RA. The ultimate aim of this project is to identify the causal genes in RA associated loci, to pinpoint disease-associated variants and to elucidate the mechanisms by which variants modify gene function.

1.6.1. Hypothesis

The causal variants within RA associated intergenic regions act by influencing gene regulation, possibly through physical contact with distal target genes and/or alteration of binding of regulatory proteins.

1.6.2. Study design

This study had the potential to determine the likely casual genes in a number of genetic loci implicated by the findings from GWAS as being associated with RA. It offered the opportunity to gain an insight into the mechanisms involved in long-range regulation of genes, and how associated variants change the interactions between regulators and promoters.

In order to address the aims of the project, a functional genomics approach utilising cutting-edge laboratory methods and bioinformatics has been used.

Stage 1: Potential causal variants may well lie some distance away from any genes, therefore chromatin folding could dictate how the variant is affecting regulatory activity. In this study, Capture Hi-C was used to characterise long-range genomic interactions involving disease associated loci.

Stage 2: Bioinformatic analysis of publicly available datasets helped to prioritise potential causal variants for further analysis and give insight into the regulatory roles of causal variants in disease associated regions.

Stage 3: The 6q23 locus was selected for in-depth study. Validation of interactions identified in the CHi-C experiments was performed using 3C-qPCR, which was also used to identify genotype-specific effects.

Stage 4: ChIP is frequently used to investigate DNA-Protein interactions occurring within the cell. It can be used to determine if a specific protein such as a transcription factor interacts with a particular genomic region. In this study, ChIP was used to investigate genome-specific binding of

markers of active regulatory regions, such as H3K4me1, H3K27ac and transcription factors in the 6q23 region, identified through bioinformatics.

Relevant cell lines implicated in autoimmunity such as B-cells and T-cells were used in the experiments. The B-cells used were HapMap EBV-transformed lymphoblastoid cell lines from individuals that matched the genotypes for the SNP being investigated. The T-cells used were an established Jurkat human leukaemic T-cell line which is commercially available and commonly used as a model T-cell line.

Ultimately, the results generated in this project gave critical insight into the behaviour of an RA associated locus, 6q23, in terms of both regulatory activity and the 3-D conformation of the genome.

2. Methods

2.1. Methods

Long-range chromatin interactions between RA associated loci and their potential targets were investigated using Capture Hi-C followed by bioinformatic prioritisation of loci for further investigation. 3C-qPCR was used to validate significant interactions in the chosen locus (6q23) and test for genotype-specific interactions, ChIP was employed to determine if the SNPs lied in regulatory regions or had evidence of altered transcription factor binding. In the first section, the general lab methods used in this study are described then the main protocols described in detail. Comprehensive tables of all the reagents, equipment and kits are included in Appendix 1 (Tables 30-33).

2.2. General lab methods

2.2.1. Cell culture

Cell lines

In order to study the effects of putative functional variants, experiments were conducted in cell types relevant to autoimmune diseases (Farh *et al.* 2015). B-cells were represented using HapMap B-lymphoblastoid cell lines (LCLs) and T-cells represented by the Jurkat E6.1 leukaemic T-lymphoblast cell line. LCLs have been genotypically well characterised as part of the HapMap project and cells carrying the three different genotypes for the variants of interest are commercially available.

a) HapMap B-Lymphoblastoid cell lines

The LCLs used in this project were obtained from Coriell (Camden, New Jersey). Cell lines were chosen according to rs6927172 genotype (HapMap and 1000 Genomes data) from individuals of European ancestry (CEPH - Utah residents with ancestry from Northern and Western Europe) (see Appendix Table 23 for cell line identifier and genotype).

LCL cultures were shipped in cell culture flasks filled to capacity with CO_2-equilibrated medium to provide sufficient nutrients for extended transport times. Upon receipt, the flasks were incubated unopened overnight at 37°C. The cultures were counted using a haemocytometer the next day and the viability checked with trypan blue stain. The cultures were either split if sufficient growth had occurred or the medium volume decreased to yield a cell density of $2x10^5$ – $5x10^5$ viable cells/ml.

LCLs grow in suspension as small (7-9µm diameter) cells which form easily dispersible aggregates. Cells were grown in vented 25cm^2 cell culture flasks containing 10-20mls of Roswell Park Memorial Institute medium + 2mM L-glutamine (RPMI-1640), supplemented with 15% foetal bovine serum (FBS). Flasks were incubated upright at 37^{o}C/5% CO_2. Cultures were regularly monitored to maintain a cell density between $2x10^5$ – $5x10^5$ viable cells/ml. Cells were split when necessary using a 1:4 split ratio into fresh medium until they reached a maximum density of $1x10^6$ cells/ml.

b) Jurkat E6.1 Human leukaemic T-lymphoblast cell line

The Jurkat cell line was established in the late 1970s from the peripheral blood of a 14 year old boy with leukaemia (Schneider *et al.* 1977). The Jurkat E6.1 clone is the standard T-cell line expressing CD4 used in immunological studies such as T-cell receptor signalling (Abraham *et al.* 2004). The cell line was obtained from LGC Standards.

Jurkat T-cells grow in suspension as small (12µm diameter) cells which may form small aggregates. 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^{o}C/5% CO_2 and the cultures regularly monitored to maintain a cell density between 3x10^5 – 9x10^5 viable cells/ml. Cells were split every two days into fresh medium until they reached a maximum density of 1x10^6 cells/ml. A 1:4-1:6 split ratio was used, as recommended by the supplier, and cells reached the desired density in approximately 4 days.

2.2.2. Analysis of DNA quantity and quality

2.2.2.1. Quant-iT™ dsDNA broad-range assay

Hi-C and 3C libraries were quantified using a Quant-iT™ double-stranded DNA broad-range (dsDNA BR) (Life Technologies) assay using a Qubit™ fluorometer. Due to large amounts of contaminants such as salt and enzymes, present in Hi-C and 3C samples due to the large reaction volumes, spectrophotometric methods such as Nanodrop are not suitable for Hi-C and 3C library quantification. The Qubit™ fluorometer utilises fluorescent dyes which only fluoresce when bound to DNA, thereby allowing accurate quantification of samples without contaminants such as salts or organic solvents also being measured. Concentration is calculated based on the relative relationship between two supplied λ dsDNA BR standards and the samples.

All reagents were equilibrated to room temperature before use. Calibration of the Qubit™ was carried out using two λ dsDNA BR standards of 0ng/µl and 100ng/µl. The Quant-iT™ dsDNA BR working solution was prepared by diluting the Quant-iT™ dsDNA BR reagent 1:200 in Quant-iT™ buffer. The assay tubes (standards and samples) were prepared in 0.5ml clear, thin-walled PCR tubes according to Table 5. Each tube was vortexed for 2-3 sec then incubated for 2 min at room temperature. The Qubit™ was calibrated with the standards then the samples analysed.

Table 5: Reaction setup for Qubit Quant-iT™ dsDNA BR assay

	Standard assay tubes	**Sample assay tubes**
Volume of working solution	190µl	180-199µl
Volume of standard (0ng/µl and 100ng/µl)	10µl	-
Volume of sample	-	1-20µl
Total volume in assay tube	200µl	200µl

2.2.2.2. Bioanalyzer assessment of DNA libraries for next-generation sequencing

The Agilent 2100 Bioanalyzer is a microfluidics-based platform that uses on-chip gel electrophoresis for the sizing, quantification and quality control of DNA, RNA and proteins. The high-sensitivity DNA-HS kit is used to analyse fragmented DNA or DNA libraries for next-generation sequencing (NGS) and was used to analyse the pre-capture and post-capture Hi-C libraries prior to sequencing on the HiSeq 2500. Quantification can be accurately performed on samples from 50-7000bp (base-pairs) in size down to 100pg/µl, and gives an accurate analysis of the fragment size range in the library. For all Bioanalyzer assessments, the chips were prepared according to the manufacturer's specific kit guidelines.

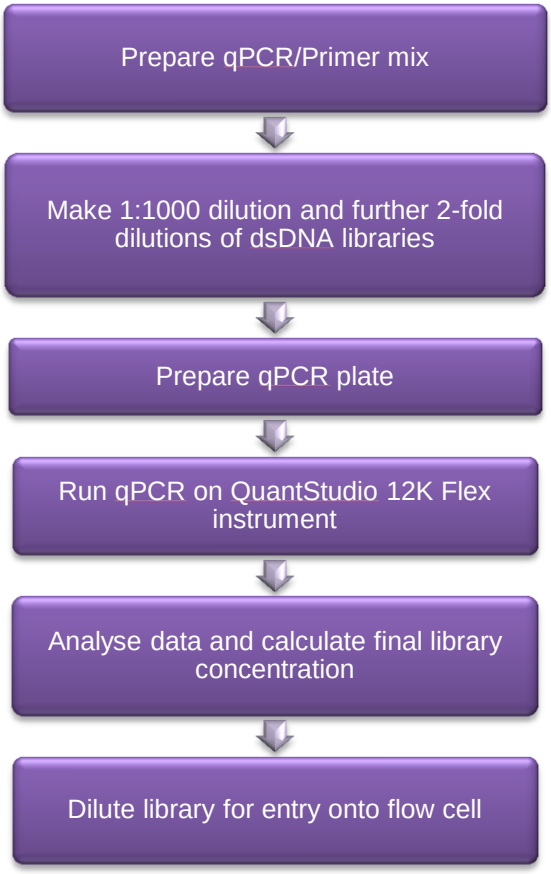
2.2.2.3. Library quantification for next-generation sequencing

The accurate quantification of DNA libraries is an essential step prior to sequencing. If the DNA concentration in the library is overestimated and the actual concentration of DNA is much lower, the cluster density on the flow cell will be too low; underestimation of a high concentration library would result in a cluster density on the flow cell that is too high. Both situations result in suboptimal sequencing. Library quantification using qPCR is regarded as the best method for accurate quantification because it counts the actual number of amplifiable molecules in the sample and the wide dynamic range allows quantification of very dilute samples (information from KAPA Biosystems).

Absolute quantification of samples is carried out by running a dilution series of known DNA concentrations, and a standard curve is generated by plotting the log of each known concentration in the dilution series (x-axis) against the C_T (threshold cycle) value for the concentration (y-axis). The standard curve can then be used to calculate the starting concentration of the unknown sample based on its C_T value. The KAPA library quantification kit (KAPA Biosystems) contains six 10-fold dilutions of KAPA 452bp Illumina standard (20-0.0002pM), 2X KAPA SYBR fast Master Mix, and 10X primer premix containing primers that are specific for libraries generated using Illumina adaptors.

The work flow is summarised in Figure 9. Firstly, 1ml of primer premix (Primer 1: 5'-AATGATACGGCGACCACCGA-3', Primer 2: 5'-CAAGCAGAAGACGGCATACGA-3'), was added to the 2X KAPA SYBR fast Master Mix and vortexed to mix. All library dilutions were made in 10mM Tris-HCl pH8.0 + 0.05% Tween 20 to reduce DNA adherence to plastics. An initial library dilution of 1:1000 (1µl DNA + 999µl dilution buffer) was prepared and vortexed to mix. Subsequent 2-fold dilutions of 1:2000, 1:4000 and 1:8000 were prepared by adding 100µl diluted library to 100µl dilution buffer, vortexing after every dilution. qPCR was carried out on a QuantStudio 12K Flex instrument (Life Technologies) using MicroAmp 384-well optical plates and covers (Life Technologies), with a 10µl reaction mix, comprising 4µl template and 6µl Master Mix.

Figure 9: KAPA qPCR library quantification for Illumina



All standards and samples were run in triplicate along with a no-template control (NTC). The following cycling parameters were used, as recommended by the kit manufacturer.

Initial activation/denaturation	95^{o}C	5 min	
Denaturation	95^{o}C	30 sec	
Annealing/extension/data acquisition	60^{o}C	45 sec	X 35 cycles

Following qPCR, the standard curve was validated by checking the correlation coefficient (R^2) value and the reaction efficiency. The samples were checked to determine if the 2-fold dilution series had C_T values spacing approximately 1 cycle apart then the concentration of each library was calculated according to the example below:

Sample and Dilution	Concentration, pM (from instrument)			Average concentration (pM)	Size adjusted concentration (pM)	Concentration of undiluted library stock (pM)
Library 1:1000	A1	A2	A3	A	A x (452/Av length) = W	W x 1000
Library 1:2000	B1	B2	B3	B	B x (452/Av length) = X	X x 2000
Library 1:4000	C1	C2	C3	C	C x (452/Av length) = Y	Y x 4000
Library 1:8000	D1	D2	D3	D	D x (452/Av length) = Z	Z x 8000

a) The calculated values for each dilution relative to the standards were obtained from the instrument.

b) A size adjustment calculation was performed to account for the difference in size between the average fragment length of the library (determined by Bioanalyzer) and the DNA standard (452bp).

c) The final concentration of the library was calculated by multiplying by the relevant dilution factor.

The average of the triplicate data points from the most concentrated library dilution that fell within the dynamic range of the standard curve was used to calculate the concentration of the undiluted library.

2.3. Investigating long-range chromatin interactions by Capture Hi-C

There is well established evidence that chromatin folding can bring genomic regions that are far apart into close proximity (Davison *et al.* 2012; Pomerantz *et al.* 2009; Zhang *et al.* 2012d) to play a role in transcriptional regulation (Fraser *et al.* 2007; Smallwood *et al.* 2013). Identifying long-range interactions between disease-associated SNPs and distal genes can give confidence that the correct gene has been identified. 3C technologies are used to study long range interactions and are discussed in section 1.4.2.

The Hi-C protocol (Lieberman-Aiden *et al.* 2009) combines 3C with NGS allowing a genome-wide view of interactions. Capture Hi-C (CHi-C) combines traditional Hi-C with a solution capture hybridisation step as shown in Figure 8 (Dryden *et al.* 2014; Jager *et al.* 2015; Mifsud *et al.* 2015). In this study a CHi-C approach was used to investigate genome-wide interactions in cell lines important in autoimmune diseases (Martin *et al.* 2015).

2.3.1. Capture Hi-C experimental design

Development of the CHi-C protocol during the early stages of my PhD made it possible to interrogate genome-wide interactions with specific target regions using Agilent SureSelect Custom Capture Libraries. Therefore, investigation of all RA, PsA, JIA and T1D loci (disease-associated regions and gene promoters) was carried out using two separate, complementary, custom-designed captures (Martin *et al.* 2015).

All independent lead disease-associated SNPs for RA were taken from both the Immunochip study (Eyre *et al.* 2012) and a transethnic GWAS meta-analysis (Okada *et al.* 2014). All RA loci which reached genome-wide significance were included in the design. Disease associated SNPs from JIA and PsA Immunochip studies were also included in the capture design (Bowes *et al.* 2015; Hinks *et al.* 2013). Using index SNPs from Immunochip fine-mapping studies increased the probability that the strongest associated variant had been identified over and above the GWAS evidence. A list of T1D SNPs, identified through credible sets analysis by a collaborator (Dr Chris Wallace, NIHR Cambridge Biomedical Research Centre, Cambridge Institute for Medical Research, University of Cambridge).

Defining credible sets is a Bayesian method used for the statistical analysis of fine-mapping data (Onengut-Gumuscu *et al.* 2015; Wallace *et al.* 2015). The evidence for association is measured by the Bayes factor which calculates the posterior probability for each SNP based on the assumption that the SNP is driving association. Identification of the causal SNP after fine-mapping can be carried out using a set of SNPs accounting for 95% or 99% of the probability. If the true causal SNP has been fine mapped it will be contained within the relevant credible SNP set.

The associated regions contained the lead SNP (the SNP showing the lowest association P value) and all SNPs in LD with the lead SNP ($r^2 \geq 0.8$) based on the 1000 Genomes Phase 1 samples of European ancestry. The LD region for al loci (including those from the transethnic GWAS meta-analysis) was defined using the European LD structure because of the larger LD block sizes. In addition, credible SNP sets were defined at 99% confidence for all RA and T1D associations (Onengut-Gumuscu *et al.* 2015) identified on the Immunochip. RA regions were extended as necessary to include the credible SNP region and any overlapping regions were merged using BEDTools v2.21.0 (Quinlan *et al.* 2010).

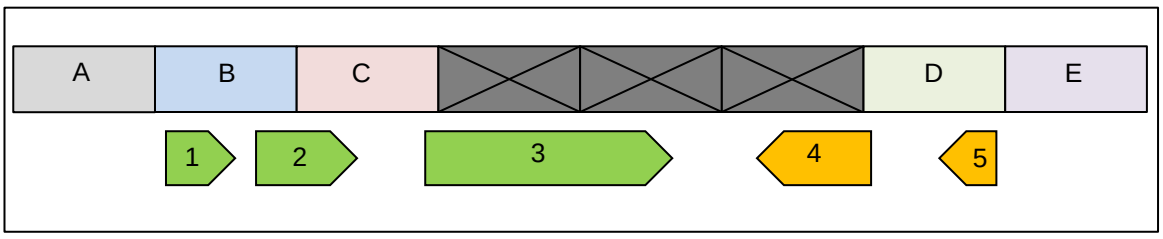
Capture probes were designed in house by the unit Bioinformatician (Mr Paul Martin) using a custom PERL script using the Ensembl release 75; GRCh37 sequence. Probe size was set at 120bp with 25-65% GC content and a maximum of 3 unknown (N) bases. Probes were only designed at restriction fragment ends within sonication size range (400bp), as close as possible to each end of the targeted *Hin*dIII fragment. Following design, the probe set was submitted to the Agilent eArray software for manufacture.

Promoter Capture

Promoter Capture target regions were defined as a 1Mb region around each disease-associated SNP. All *Hin*dIII restriction fragments within 500bp 5' of the transcription start site of all Ensembl Release 75; GRCh37 gene transcripts within the defined region were targeted. A positive control region containing the well-characterised *HBA* (Haemoglobin A) locus was also included (Hughes *et al.* 2014; Schoenfelder *et al.* 2010b).

The schematic shown below in Figure 10 shows how the regions included in the promoter capture were designed. Fragments defined by *Hin*dIII restriction sites are shown by rectangles and the targeted fragments are labelled A-E. Untargeted fragments are crossed out. The gene transcripts are shown by arrows, labelled 1-5, with the arrow head pointing in the direction of transcription. The fragments would be targeted for each transcript as follows: 1 – A & B; 2 – B; 3 – C; 4 – D; 5 – D & E. Transcripts 1 and 5 have two fragments to target as the distance from the transcription start site to the next restriction site is less than 500bp.

Figure 10: Schematic showing promoter capture design

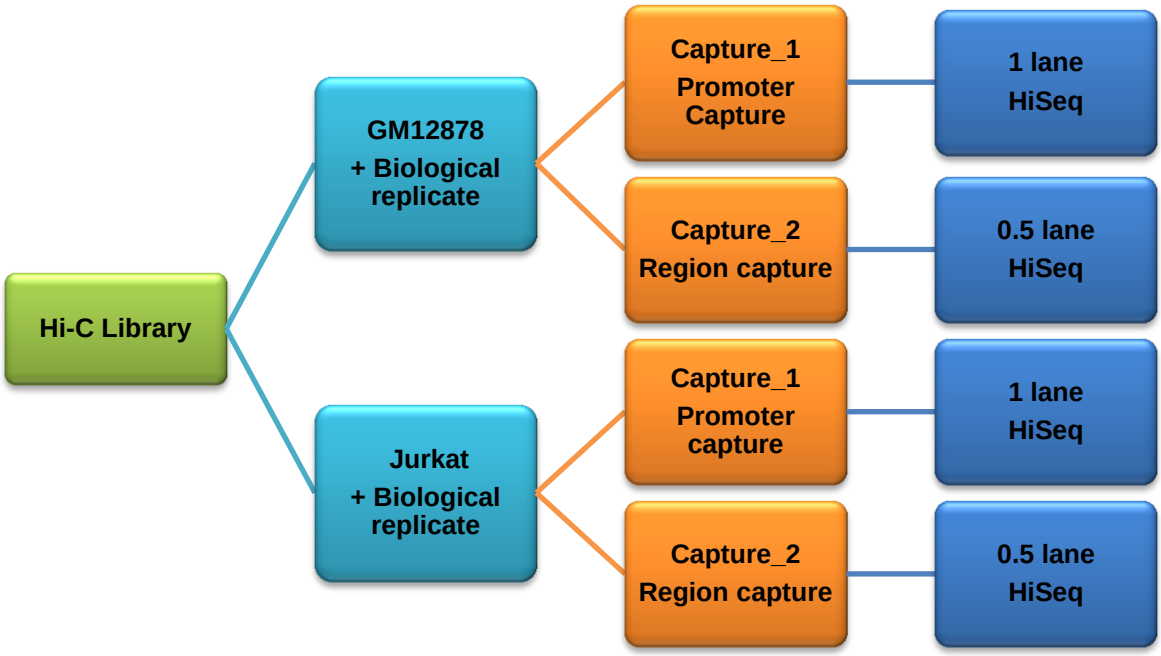


Region Capture

All *Hind*III restriction fragments within the disease associated region, as defined by r^2>0.8 or credible SNP sets, not containing a gene promoter and therefore already included in the Promoter Capture experiment, were targeted. If there was <500bp between the region start/end and the restriction site the region was extended by one restriction fragment. A positive control region containing the *HBA* locus was also included (Hughes *et al.* 2014; Schoenfelder *et al.* 2010b).

The experimental plan is summarised in Figure 11. Briefly, duplicate Hi-C libraries were prepared for GM12878 and Jurkat cell lines. The same library was used to conduct both Promoter Capture and Region Capture experiments. Solution capture hybridisation was carried out using Agilent SureSelectXT reagents and protocol then the libraries sequenced on an Illumina HiSeq 2500.

Figure 11: Capture Hi-C Experimental plan

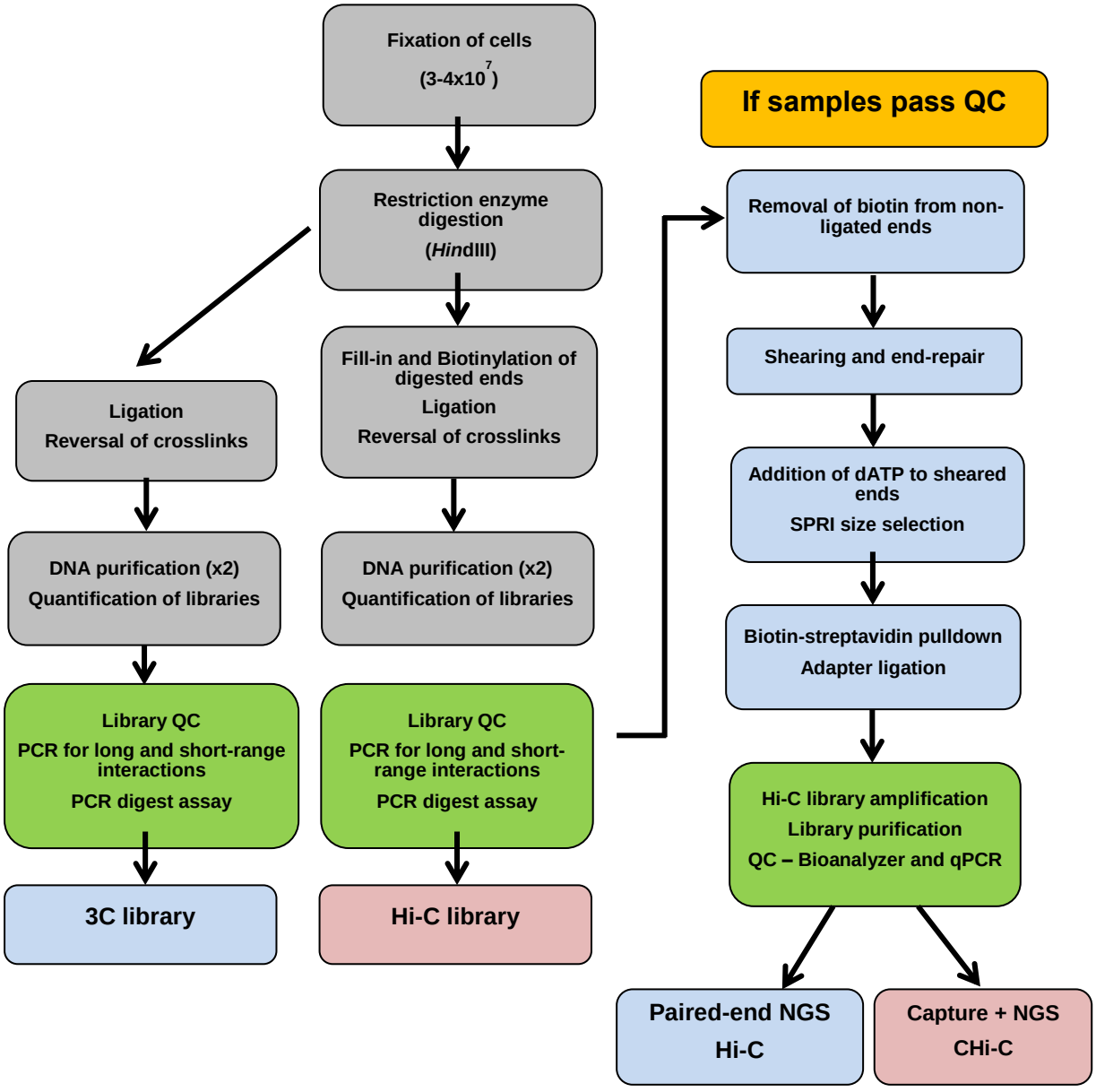


2.3.2. Generation of Hi-C libraries

The Hi-C protocol (Belton *et al.* 2012; Duan *et al.* 2012; Lieberman-Aiden *et al.* 2009) is based on 3C (Dekker *et al.* 2002), a well-established technique used to investigate long-range chromatin interactions (Figure 7). Hi-C can be used to show the overall genome structure, the biophysical properties of chromatin and long-range interactions between genes and regulatory elements. It is this final function that is of particular interest to this project.

Briefly, Hi-C involves the purification of 3C ligation products which have been biotinylated to specifically identify ligation junctions through NGS. The steps involved are summarised in Figure 12. The results of Hi-C can show chromosomal interactions across the entire genome in an 'all-to-all' manner as opposed to 3C which is 'one-to-one'.

Figure 12: Generation of Hi-C libraries



Hi-C Detailed protocol

Step 1: Crosslinking of DNA-protein interactions

Cells were grown to ~90% confluence and counted using a CASY automated cell counter. Approximately $3x10^7$ cells were prepared from each cell line (duplicate samples were prepared to give ~$5\text{-}6x10^7$ cells per experiment). Cell suspensions were made up to 40ml volume with room temperature Dulbecco's modified eagle medium (DMEM).

Crosslinking of DNA-protein interactions was carried out by the addition of formaldehyde (FA) (37% stock solution) to a final concentration of 2% and the cells fixed for exactly 10 min at room temperature whilst mixing on a rocker. The crosslinking reaction was quenched by the addition of cold 1M glycine (0.125M final) followed by incubation for 5 min at room temperature, then 15 min on ice. The cells were centrifuged at 450 *x g* for 10 min at 4°C, the supernatant discarded and the pellet carefully resuspended in cold PBS (phosphate buffered saline) in a final volume of 50ml. The cells were re-centrifuged, the supernatant discarded and the cells snap-frozen in liquid nitrogen, then stored at -80°C.

Step 2: Lysis of cells and *Hin*dIII digestion

Cells were thawed on ice and resuspended in freshly prepared ice-cold lysis buffer (10mM Tris-HCl pH 8, 10mM NaCl, 0.2% Igepal CA-630, one protease inhibitor cocktail tablet). Routinely, two pellets from each cell line were resuspended and combined in 7ml complete lysis buffer to give ~$5\text{-}6x10^7$ cells. Cells were lysed for a total of 30 min as detailed below.

The cell suspensions were left on ice with occasional mixing for 10 min then transferred to a sterile Dounce homogeniser. The cells were lysed by 10 strokes of the homogeniser, a 5 min rest then 10 more strokes. After the final strokes, the cells were transferred to the remaining lysis buffer for the remainder of the 30 min incubation (now lysing in 50ml volume). Following lysis, the nuclei were collected by centrifugation at 650 *x g* for 5 min at 4°C and the supernatant discarded. To remove any remaining lysis buffer 1.25x NEBuffer2 (NEB) was layered on top of the pellet, without resuspending the nuclei, and the buffer discarded. Each pellet was resuspended in 2ml NEBuffer2 to give 8 x 250µl aliquots, each containing nuclei from $5\text{-}6x10^6$ cells – 2 aliquots were used to make a 3C control library and the remaining 6 aliquots used for Hi-C. To each of the aliquots, 108µl 1.25x NEBuffer2 was added to make a final volume of 358µl.

To remove proteins not directly crosslinked to DNA 11µl 10% SDS was added to each of the aliquots, mixed carefully, and the samples incubated at 37°C for 60 min, rotating at 950rpm. The SDS was quenched by adding 75µl 10% Triton X-100 (Sigma Aldrich) to each sample, and the samples incubated at 37°C for 60 min, rotating at 950rpm.

The chromatin was digested by the addition of 1500 units *Hin*dIII (NEB) per tube then incubated at 37°C overnight while rotating at 950rpm.

Step 3: Biotinylation of DNA ends and ligation

3C technologies are based on proximity ligation of DNA fragments under dilute conditions. A large volume and low concentration of DNA strongly favours the formation of ligation products within a single molecule (intra-molecular = *cis* interactions) rather than between two molecules (inter-molecular interactions = *trans*). Following restriction digestion and ligation, DNA fragments within the same chromatin complex behave as a single molecule that is joined together at a restriction site (Lieberman-Aiden *et al.* 2009). Biotinylation of DNA ends prior to ligation marks the junction where the DNA fragments are ligated together allowing for selection of ligated fragments using a streptavidin pull-down.

In order to improve ligation reaction efficiency and reduce experimental noise a variation of the Hi-C protocol was utilised, which uses in-nucleus ligation as opposed to in-solution ligation (Nagano *et al.* 2013; Nagano *et al.* 2015). In this protocol, the nuclei are not lysed with SDS prior to ligation meaning ligation is carried out inside the nucleus.

Before starting, a 15ml tube was prepared for each sample containing 6.71ml water and 82µl BSA (10mg/ml) (ligation buffer) and kept on ice until needed. The 3C control aliquots were kept at 37°C during this step as they are not biotinylated. To fill in the restriction fragment overhangs and mark the DNA ends with biotin, 6µl 10xNEBuffer2, 2µl H_2O, 1.5µl 10mM dCTP, 1.5µl 10mM dGTP, 1.5µl 10mM dTTP, 37.5µl 0.4mM biotin-14-dATP (Life Technologies), and 10µl 5U/µl Klenow (DNA polymerase I large fragment – NEB) was added to each Hi-C aliquot. The samples were carefully mixed and incubated for 60 min at 37°C without rotation, with resuspension of the nuclei every 10 min.

When the biotinylation/fill-in reaction was completed, the samples were placed on ice. To each 15ml tube containing ligation buffer 820µl 10xT4 DNA ligase buffer (NEB) was added then each sample (both Hi-C and 3C) transferred to the appropriate 15ml tube and thoroughly mixed to collect all nuclei. To the Hi-C aliquots 50µl 1U/µl T4 DNA ligase (Life Technologies) was added, and to the 3C aliquots 2µl T4 DNA ligase (NEB) was added. The lids of the tubes were tightly closed, the samples mixed and the samples incubated for 4-6 hours at 16°C. Following ligation, crosslinks were reversed and protein degraded by the addition of 60µl 10mg/ml proteinase K (Roche) per tube and overnight incubation at 65°C.

Step 4: DNA purification part I

The large reaction volume used in the ligation reactions means that the samples contain many contaminants such as salts from the buffers and excess enzymes. DNA purification is carried out by organic solvent extraction followed by ethanol precipitation, firstly in a large volume, and then a second extraction using a smaller volume.

Following overnight incubation, an additional 60µl 10mg/ml proteinase K was added per tube and the incubation continued at 65°C for a further 2 hours. The reaction mixtures were cooled to room

temperature, then 12.5µl 10mg/ml RNaseA (Roche) added to each sample and incubated at 37°C for 60 min.

The reaction mixtures were transferred to 50ml Phase-Lock Gel (PLG) Light tubes (5-Prime) then 8ml phenol pH8.0 (Sigma Aldrich) added to each sample. The samples were mixed well for 1 min then centrifuged for 10 min at 1500 *x g*. Following centrifugation 2ml 1xTE buffer (10mM Tris-HCl pH 8.1, 1mM EDTA) was added to each sample to increase the volume to 10ml. The extraction was repeated in the same PLG tube using 10ml phenol pH8.0:chloroform:isoamyl alcohol (Sigma Aldrich), following the steps described previously. After centrifugation, the supernatant was transferred to a fresh 50ml tube. To precipitate the DNA, 1/10 volume of 3M sodium acetate pH5.2 (Lonza) and 2.5 volumes of ice cold 100 % ethanol was added to each tube then incubated overnight at -20°C.

Step 5: DNA purification part II

Following overnight precipitation, the samples were centrifuged in a microfuge at 2500 *x g* at 4°C for 30 min. The supernatant was removed then the pellets dried for 45-60 min in a 37^{o}C oven. Each pellet was resuspended in 400µl 1xTE then purified by two phenol pH8.0:chloroform:isoamyl alcohol extractions in 2ml PLG light tubes.

To each sample 400µl phenol pH8.0:chloroform:isoamyl alcohol (1:1) was added and the samples mixed for 1 min. The tubes were centrifuged for 10 min at 14,500 *x g* then a second extraction carried out in the same PLG tube. After centrifugation, the supernatant was transferred to a fresh 2ml tube and the DNA precipitated as previously described in Step 4.

Step 6: DNA purification part III

The precipitated DNA was centrifuged at 14,500 *x g* at 4°C for 30 min then each pellet washed three times with 70% ethanol. After the final wash, all of the ethanol was carefully removed then the pellets dried at 37^{o}C for no more than 5 min. Each DNA pellet was resuspended fully in 25µl 1xTE buffer and the contents of the Hi-C tubes pooled (3C samples were also pooled but kept separate). Hi-C and 3C libraries were quantified using a Quant-iT™ dsDNA BR assay using a Qubit™ fluorometer (See section 2.2.2.1).

Step 7: Hi-C ligation efficiency and quality controls

Quality control is an essential part of the Hi-C protocol to ensure overall library quality and successful incorporation of the biotin label at the ligation junction. Agarose gel electrophoresis can give a good indication of library quality and PCR should be carried out to detect known short-range and long-range interactions. The successful fill-in and ligation of a *Hin*dIII site (AAGCTT) creates a site for the restriction enzyme *Nhe*I (GCTAGC) (Lieberman-Aiden *et al.* 2009), which can be detected by a PCR digest assay. The 3C amplicon should only digest with *Hin*dIII, not *Nhe*I, whilst the Hi-C amplicon should digest with *Nhe*I, not *Hin*dIII, if the fill-in and biotinylation reactions have been successful.

a) Agarose gel electrophoresis

To check the quality and quantity of the libraries, 2µl and 6µl aliquots of 1:10 dilutions from the Hi-C and 3C libraries were electrophoresed at 120V for 50 min on a 0.8% agarose gel stained with ethidium bromide then visualised on a transilluminator.

b) PCR to detect short-range and long-range interactions

PCRs to detect previously described short range (Belton *et al.* 2012; Lieberman-Aiden *et al.* 2009) and long range interactions were carried out using a range of primers (Appendix Table 26). Primers were diluted 1:10 from a 100µM stock to make a 10µM working stock. Template DNA was normalised to 200ng and 1µl used per reaction.

A stock solution of 5X PCR mix was prepared to use in the PCR reaction (500µl HotStar 10X PCR mix, 10µl each 10mM dNTP, 460µl water). PCR reactions were set up using either 96-well plates or 8-well strips with domed caps according to the recipe below.

5 x PCR mix	5µl
Each primer	1µl + 1µl
Template/water	1µl
Polymerase (HotStar – Qiagen)	0.5µl
Water	16.5µl

PCR was performed using the following cycling parameters on a BioRad T100 thermocycler (or equivalent).

95^{o}C	15 min	
60^{o}C	1 min	x36
72^{o}C	1 min	
94^{o}C	30 sec	
60^{o}C	2 min	
72^{o}C	10 min	
4^{o}C	Hold	

Following PCR, amplification products were electrophoresed at 120V for 50 min on a 1.5% agarose gel stained with ethidium bromide then visualised on a transilluminator.

c) Verification of Hi-C marking and Hi-C ligation efficiency by PCR digest assay

Five (identical) 25µl PCR reactions using 200ng of library per reaction were set up to amplify a short-range ligation product formed from two nearby restriction fragments, using the *Hin*dIII Dekker (Lieberman-Aiden *et al.* 2009) or AHF control primers (Belton *et al.* 2012) and region of interest

primers spanning adjacent *Hin*dIII sites. Following amplification using the parameters above, the PCR products were pooled, purified using a Qiagen PCR purification kit according to the manufacturer's protocol, and the concentration determined by Nanodrop.

Subsequently, the purified pooled samples were split into four samples: undigested, digested with 1µl *Hin*dIII, digested with 1µl *Nhe*I, and digested with 1µl both *Hin*dIII and *Nhe*I. Reactions were carried out using 500-600ng of PCR product per digest in 20µl reaction volumes containing 1xNEB CutSmart buffer with the volume adjusted with water for 1.5-2 hours at 37^{o}C. Following digestion, the samples were electrophoresed at 120V for 75 min on a 1.5% agarose gel stained with ethidium bromide then visualised on a transilluminator. If samples passed QC the protocol was continued.

Step 8: Removal of biotin from non-ligated DNA ends

Removal of biotin from non-ligated ends is carried out by the action of T4 DNA polymerase, which removes nucleotides from unligated ends through 3'-to-5' exonuclease activity. Replicate 5µg aliquots of Hi-C library, up to a total of 40µg, were mixed with 0.5µl 10mg/ml BSA, 5µl 10x NEBuffer2, 2µl 2.5mM dATP, and 5µl T4 DNA polymerase in a total volume of 50µl then incubated at 20°C for 4 hours. The reactions were stopped by adding 2µl 0.5M EDTA pH 8.0 to each tube and two reactions pooled to give a total amount of ~10µg DNA per sample.

To purify the DNA, a single 1:1 phenol pH8.0:chloroform:isoamyl alcohol extraction was carried out using 2ml PLG Light tubes as previously described. After centrifugation, the supernatant was transferred to a fresh 2ml tube and the DNA precipitated as previously described.

Following precipitation the samples were centrifuged at 14,500 *x g* for 30 min at 4^{o}C. The supernatant was removed and the pellets washed twice with 1ml fresh 70% ethanol by centrifuging at 14,500 *x g* for 10 min at 4^{o}C. After the final wash, all the ethanol was removed and the pellets dried for no longer than 5 min. To resuspend the pellets, 130µl water was added to each tube and the samples either stored at -20^{o}C or continued to the next step.

Step 9: DNA shearing and end repair

For Hi-C library shearing, a Covaris S220 was used. Following DNA shearing any 5'-overhangs are filled in by the action of T4 DNA polymerase and 3'-overhangs removed by the action of the Klenow enzyme. T4 polynucleotide kinase is used in the end-repair reaction to add a 5' phosphate group to the Hi-C library, which enables the ligation of sequencing adapters to the Hi-C libraries in subsequent steps.

To shear the DNA, each 130µl sample from Step 8 was transferred to a Covaris Microtube (the maximum volume) and the following parameters used to obtain fragments with a peak around 400bp:

Duty factor	10%
Peak Incident Power (W)	140
Cycles per burst	200
Time	55 sec

After shearing, the entire volume of each sample (130µl) was transferred into a fresh 1.5ml tube, the following reagents added to each sample and the samples incubated for approximately 30 min at room temperature to repair the sheared ends:

10x ligation buffer	18µl
2.5mM dNTP mix	18µl
T4 DNA polymerase	6.5µl
T4 DNA polynucleotide kinase	6.5µl
Klenow (Large)	1.3µl

Following end-repair, each sample was split into two, each containing ~5µg DNA, and purified using a modified Qiagen MinElute protocol (Belton *et al.* 2012).

- Five volumes buffer PB was added to each sample
- Samples were centrifuged at 6000 *x g* for 1 min then 16,000 *x g* for 1 min, and the flow through discarded
- Samples were washed with 750µl buffer PE by centrifuging at 16,000 *x g* for 1 min
- The flow through was discarded and the small droplet of PE that was resting on the lip, above the membrane was removed with a pipette
- Samples were centrifuged at 16,000 *x g* for 1 min then the column was placed in a fresh 1.7ml tube
- To elute the DNA 20µl HOT (65^{o}C) TLE (made fresh; 10mM Tris, 0.1mM EDTA) was added to the column and incubated at room temperature for 2 min
- Samples were centrifuged at 6000 *x g* for 1 min then 16,000 *x g* for 1 min
- The elution was repeated with 15µl HOT TLE and the samples transferred to fresh tubes

Step 10: Addition of dATP and size selection

Klenow 3'→5' exo^{-} lacks exonuclease activity but retains 5'-3' polymerase activity. The A-tailing reaction adenylates the 3' end, allowing the ligation of sequencing adapters to the Hi-C libraries.

a) Addition of dATP

To perform the A-tailing reaction the following reagents were added to the sheared, end repaired DNA from Step 9 (30µl) and the reactions incubated at 37°C for 30 min.

10 x NEBuffer 2	5µl
1mM dATP	11.5µl
Klenow exo-	1µl

To inactivate the enzyme, the samples were incubated at 65°C for approximately 20 min, then put on ice immediately afterwards.

b) SPRI size selection

Traditionally, size selection has been carried out by excising DNA fragments from agarose gel. The use of silica columns or magnetic beads for DNA clean-up is faster, more controllable and removes the need for ethidium bromide. Solid phase reversible immobilisation (SPRI) is a magnetic bead DNA clean-up that is specific for double-stranded DNA (DeAngelis *et al.* 1995). Polystyrene beads are coated in magnetite and carboxyl molecules that reversibly bind to DNA in the presence of a solution of 20% polyethylene glycol and 2.5M NaCl. Double-sided SPRI size selection can be used to remove small products such as primer dimers and larger fragments so that the size range can be tightly controlled. A low SPRI:DNA ratio eg. 0.7X binds large products and the smaller products remain in the supernatant. A higher ratio of SPRI:DNA eg. 1.0X can then be used to bind the fragments of the correct size, leaving the small DNA in solution. The correctly sized DNA can then be eluted from the beads.

Ampure XP beads are paramagnetic beads used for SPRI purification of PCR products (DeAngelis *et al.* 1995) and size selection of DNA libraries for NGS. The beads are suspended in an optimised buffer that selectively binds PCR amplicons of ≥100bp to the beads. Excess primers, dNTPs, salts and enzymes are removed by washing with 70% ethanol, resulting in a pure DNA product.

DNA fragments between 200-650bp were selected by double-sided SPRI bead size selection (0.6x followed by 0.9x). Ampure XP SPRI beads were thoroughly mixed by vortexing then allowed to equilibrate at room temperature for at least 30 min before use.

Two A-tailed samples were pooled into a fresh tube (A) (total volume now 100µl). For each sample, one tube (B) with 180µl SPRI bead solution was prepared. 60µl of SPRI bead solution from tube B was added to tube A containing 100µl of DNA solution (0.6x). The samples were thoroughly mixed, incubated for 10 min at room temperature, then placed on a magnet. The unbound supernatant containing the DNA in the desired size range was recovered into a fresh tube (tube C). Tube A containing the beads was discarded.

The SPRI beads were concentrated by placing tube B (containing 120µl of SPRI beads in solution) on the magnet, and removing all but 30µl of the supernatant (i.e. approximately 90µl of the supernatant was discarded). The beads in tube B were resuspended in the remaining 30µl volume. Concentrated beads (30µl) from tube B was added into tube C (0.9x SPRI bead, i.e. ratio of DNA to

SPRI beads in solution is now 1:0.9). The samples were mixed well, incubated for at least 10 min at room temperature, placed on the magnet, and the supernatant discarded.

The beads were washed twice with freshly prepared 70% ethanol, leaving the samples on the magnet. Tube B was discarded. Bead-bound DNA in tube C was resuspended in 50µl TLE, incubated at room temperature for 5 min, placed on the magnet and the supernatant (containing size-selected DNA) transferred into a fresh tube D. Tube C containing the beads was discarded.

All the Hi-C library aliquots were pooled and dilutions of each library prepared (1:20, 1:50 and 1:100 dilutions). Libraries were quantified by Qubit™ using the Quant-iT™ dsDNA BR assay, as described previously, with an expected yield from 40µg starting material of ~10µg.

Step 11: Adapter ligation and Biotin-streptavidin pulldown

For paired-end sequencing, each sample needs specific adapters ligated to the ends, one of which is Universal and contains the sequence that attaches the sample to the flow cell. The other adapter contains the barcode, which is a set of 6 nucleotides in the middle of the sequence used to identify the sample. Barcoding gives each sample a unique identity, allowing multiple pooled samples to be run on the same sequencing lane (Illumina MiSeq or HiSeq for these experiments), reducing the cost of sequencing. Illumina uses a green laser to sequence G/T nucleotides and a red laser to sequence A/C nucleotides. At each cycle at least one of two nucleotides for each colour channel needs to be read. It is important to maintain colour balance for each base of the index read being sequenced, otherwise the index read sequencing could fail.

a) Adapter ligation

Hi-C libraries for promoter/region capture were prepared using short TruSeq adapters and the Indexes added following solution capture hybridisation (See Appendix Table 27 for sequences).

Adapters were generated by annealing TruPE_adapter_1 with TruPE_adapter_2. To anneal the adapters in a thermoblock: 15µM TruPE_adapter_1 + 15µM TruPE_adapter_2 were mixed in equal amounts in a 1.5ml tube then incubated at 95^{o}C for 15 min. The temperature was reset to 70^{o}C and when it reached 70^{o}C, timed for 15 min. The thermoblock was reset to 22^{o}C and the samples allowed to slowly cool down. Adapter aliquots (15µM) of 10-20µl were made and stored at -20°C, and an aliquot thawed just before use.

b) Biotin-streptavidin pulldown

Dynabeads® MyOne™ Streptavidin C1 beads are magnetic beads pre-coupled with a streptavidin ligand which has extremely high affinity for biotin (Dechancie *et al.* 2007). During incubation, the biotinylated sample binds to the beads and the complex is captured on a magnet. The unbound material (non-biotinylated products) can be removed by aspiration and the bead-bound target washed. The bead-bound target can either be eluted, or in the case of this experiment, used directly whilst attached to the beads.

An excess of wash buffers were prepared for use in the biotin-streptavidin pulldown:

a) TB (Tween buffer) 1M NaCl, 5mM Tris-HCl pH8.0, 0.5mM EDTA, 0.05% Tween - 1600µl per sample

b) 1 x NTB (No tween buffer) 1M NaCl, 5mM Tris-HCl pH8.0, 0.5mM EDTA - 600µl per sample

c) 2 x NTB (2xNo tween buffer) 2M NaCl, 10mM Tris-HCl pH8.0, 1mM EDTA - 300µl per sample

d) 1 x T4 DNA ligase buffer - 150µl per sample

e) 1 x NEBuffer2 - 300µl per sample

For the pulldown, 150µl of Streptavidin C1 beads suspension was transferred into a lo-bind 1.5ml tube. One reaction was set up per 2µg to 2.5µg of DNA as determined after SPRI size selection. Beads were washed twice with 400µl TB using the following steps for every wash:

a) Sample placed on magnetic separator and beads reclaimed

b) Supernatant discarded

c) New buffer added, thoroughly mixed, and sample transferred to a new tube

d) Sample rotated for 3 min at room temperature

e) See a)

Beads were resuspended in 300µl of 2xNTB. If the amount of Hi-C library DNA exceeded 2.5µg, the appropriate number of Hi-C samples, each containing a maximum of 2.5µg of DNA, in a total volume of 300µl TLE was prepared. The beads were combined with the Hi-C DNA making a total volume of 600µl and rotated slowly for 30 min at room temperature.

The samples were placed on a magnetic separator, beads with bound Hi-C DNA captured, supernatant discarded and the beads washed once with 400µl 1xNTB, followed by another wash with 200µl 1x ligation buffer. The beads were resuspended in 50µl 1x ligation buffer and transferred to a fresh tube. To each sample 4µl of annealed 15µM adapter and 4µl of T4 DNA ligase was added and the samples rotated slowly at room temperature for 2 hours.

Samples were placed on the magnet, Hi-C bound beads reclaimed and washed twice with 400µl TB, 200µl 1xNTB, then 200µl 1xNEBuffer2. Finally, the samples were washed with 60µl 1xNEBuffer 2, the beads resuspended in 40µl 1xNEBuffer2 and transferred into a fresh tube. If more than one streptavidin-biotin pulldown per Hi-C library was performed (i.e. if the starting amount of Hi-C library exceeded 2.5µg DNA), the reactions were pooled and stored at 4°C.

Step 12: Test PCRs to determine conditions for Hi-C library amplification

To determine the optimal number of PCR cycles for Hi-C library amplification, test PCRs were set up with n (6, 7, 9, and 12) amplification cycles. See Appendix Table 27 for details of primer sequences.

Four reactions were prepared, each containing:

Hi-C library DNA on beads	2.5µl
Buffer 5x (Phusion NEB F531)	5µl
dATP 10mM	0.7µl
dCTP 10mM	0.7µl
dGTP 10mM	0.7µl
dTTP 10mM	0.7µl
TruPE_PCR_1.0.33	0.075µl of 100µM stock
TruPE_PCR_2.0.33	0.075µl of 100µM stock
Phusion polymerase	0.3µl
H_2O	14.25µl

Four separate PCRs were carried out with the following conditions for n (6, 7, 9, and 12) cycles:

98°C	30 sec	x1 cycle
65°C	30 sec	
72°C	30 sec	
98°C	10 sec	n-2 cycles
65°C	30 sec	
72°C	30 sec	
98°C	10 sec	x1 cycle
65°C	30 sec	
72°C	7 min	
4°C	Hold	

The amount of amplified DNA was visualised by running the entire reaction (25µl) on a 1.5% agarose gel and the number of cycles for the final amplification determined from the gel.

Step 13: Final PCR amplification of Hi-C libraries

To ensure the complexity of the library and reduce the chance of artefacts and duplication sequences being introduced by PCR, the number of cycles was kept to a minimum and the PCRs carried out in many replicates which were pooled post-PCR and cleaned up using Ampure XP beads. The eluted DNA was the final Hi-C library which could either be sequenced (Hi-C) or carried forward into target capture experiments (CHi-C).

Multiple PCR reactions were set up (the remaining volume of Hi-C library divided by factor 2.5) with 25µl each, as described above, with one PCR condition determined from the test PCR (i.e. number of cycles). For both the Promoter Capture and Region Capture experiments, for both cell lines and biological replicates, the volume of beads remaining after test amplifications was split into two so that both captures came from the same Hi-C library. Following PCR, all individual PCR reactions were pooled and purified using SPRI beads.

The pooled reactions were placed on a magnetic separator, and the supernatant transferred into a fresh 1.5ml lo-bind tube. The streptavidin beads were resuspended in the original amount of 1xNEBuffer2 and kept as a backup. The volume of the supernatant containing the amplified Hi-C library was determined and purified by adding 1.8 x volumes of SPRI beads and incubating at room temperature for 10-20 min. Beads were captured on a magnet, the supernatant discarded and the beads washed twice in freshly prepared 70% ethanol. The beads were dried for 3 min at 37^{o}C then resuspended in 100µl of TLE, incubated at room temperature for 5 min, the beads collected on a magnetic separator and the supernatant (100µl) transferred to a clean lo-bind tube.

The SPRI bead purification was repeated by adding 180µl of SPRI beads to 100µl of Hi-C library. Beads were resuspended in a final volume of 25µl TLE, incubated at room temperature for 5 min, beads captured on a magnetic separator and the supernatant (~23µl to prevent bead carry-over) transferred to a clean lo-bind tube. The quality and quantity of the Hi-C library was analysed on a Bioanalyzer HS-DNA chip and by qPCR (KAPA library quantification for Illumina) (section 2.2.2.3).

2.3.3. Solution Capture Hybridisation

Step 1: Hybridisation reaction

Hybridisation of SureSelect custom capture libraries (see Appendix Table 33 for details) to Hi-C libraries was carried out using Agilent SureSelectXT reagents and protocols. Hi-C samples (400-750ng, depending on library concentration) were concentrated in a vacuum concentrator (Eppendorf) at 30^{o}C for 20-30 min, depending on the sample volume, then resuspended in 3.4µl water and kept on ice until needed. Hybridisation buffer was prepared as detailed below and warmed at 65^{o}C for 5 min before use to dissolve precipitate then 40µl per sample was added to an 8-well strip and kept at room temperature.

Reagent	**For 2 captures**
SureSelect Hyb #1 (orange cap)	50µl
SureSelect Hyb #2 (red cap)	2µl
SureSelect Hyb #3 (yellow cap)	20µl
SureSelect Hyb #4 (black cap)	26µl
Total	98µl (40µl per sample)

The SureSelect capture library mix was prepared in an 8-well strip on ice as follows:
RNase block dilution (1:9) (purple cap) (5µl per reaction + excess) was prepared by mixing 2µl RNase Block with 18µl water. For each sample:

SureSelect library 2µl

RNase Block dilution 5µl

The SureSelect Block mix was prepared as follows:

Reagent	**For 2 captures**
SureSelect Indexing block #1 (green cap)	5µl
SureSelect Block #2 (blue cap)	5µl
SureSelect Indexing block #3 (brown cap)	1.2µl
Total	11.2µl

DNA library was prepared in an 8-well strip by adding 3.4µl library to an 8-well strip then 5.6µl SureSelect Block mix was added to each sample and mixed well. The strip was sealed with a cap and the hybridisation reaction started.

Hybridisation reaction

The PCR machine (BioRad T100) was set to the following program (95˚C for 5 min then 65˚C forever) using the heated lid at 105°C throughout.

The PCR strip containing the pond Hi-C library was transferred to the PCR machine, in the position marked in red below, and the PCR program started.

A											
B											
C											
D											DNA
E											
F											
G											
H											

After just over 5 min (once the temperature reached 65˚C) the PCR strip containing hybridisation buffer was transferred to the PCR machine, in the position marked in blue and incubated for 5 min.

A											
B											Hyb
C											
D											DNA
E											
F											
G											
H											

After 5 min (10 min since the start of the PCR program), the PCR strip with the biotinylated RNA bait was transferred to the PCR machine, in the position marked in green and incubated for 2 min.

A											
B											Hyb
C											
D											DNA
E											
F											RNA
G											
H											

After 2 min, the lids were removed from the PCR strips containing the hybridisation buffer and the biotinylated RNA bait. 13µl of hybridisation buffer was pipetted into the 7µl of RNA bait (blue into green). The PCR strip containing the hybridisation buffer was discarded and the next step carried out immediately.

A											
B											
C											
D											DNA
E											
F											RNA + Hyb
G											
H											

The lid from the PCR strip containing the pond Hi-C library (DNA) was removed. 10µl of the Hi-C library was pipetted into the 20µl of RNA bait/hybridisation buffer (red into green). The empty PCR strip that contained the Hi-C library was discarded.

A											
B											
C											
D											
E											
F											RNA + Hyb + DNA
G											
H											

The remaining PCR strip (now containing Hi-C library/hybridisation buffer/RNA bait) was closed with a PCR strip tube lid immediately and incubated for 24 hours at 65˚C.

Step 2: Streptavidin-Biotin pulldown and washes

Before use, DynaBeads T1 were vortexed to mix, then 50µl beads per sample added to a 1.5ml tube and the following wash steps carried out:

a) Add 200µl SureSelect binding buffer (BB) and transfer to a fresh tube

b) Mix on vortex (low to medium setting) for 5 sec

c) Reclaim beads on magnetic separator and discard supernatant.

Steps a-c were repeated for a total of 3 washes. After the final wash the beads were transferred to a fresh tube leaving beads in 200µl BB in a fresh tube for each sample.

Biotin-streptavidin pulldown

The PCR machine was opened with the program still running (after 24 hours hybridisation) and the entire reaction transferred into the tube containing beads + BB. Samples were thoroughly mixed and incubated on a rotator for 30 min at room temperature (speed = 7).

After 30 min the beads were reclaimed and supernatant discarded. Beads were resuspended in 500µl wash buffer 1 (WBI), transferred to a fresh tube and incubated at room temperature for 15 min, vortexing every 2-3 min for 5 secs. After 15 min the beads were reclaimed and supernatant discarded. The beads were resuspended in 500µl WBII, transferred to a fresh tube and incubated at 65^{o} for 10 min, vortexing every 2-3 min for 5 sec. After 10 min the beads were reclaimed and supernatant discarded. The wash buffer II (WBII) wash was carried out a further two times for a total of 3 washes in WBII. After removing supernatant after the final WBII wash the beads were resuspended in 200µl 1xNEBuffer2, immediately transferred to a fresh tube and the beads reclaimed. The beads containing the RNA/DNA 'catch' were resuspended in 30µl 1xNEBuffer2 transferred to a fresh tube and stored at 4^{o}C.

Step 3: Determining optimum amplification

To determine the optimal number of PCR cycles for library amplification, test PCRs were set up with 7, 9, and 12 amplification cycles using the conditions detailed in section 2.3.2 except using post-capture primers. The primers used in post-capture amplification were designed to add the barcoding index to allow multiplex sequencing (Table 6).

Table 6: Final amplification PCR index sequences

Sample name	Experiment	Final amplification PCR primers	Index	Adapter sequence
GM12878_ProCap	Promoter Capture	Universal + Indexed Primer	AR006	GCCAAT (Added post-capture)
Jurkat_ProCap	Promoter capture	Universal + Indexed Primer	AR012	CTTGTA (Added post-capture)
GM12878_RegCap	Region capture	Universal + Indexed Primer	AR003	TTAGGC (Added post-capture)
Jurkat_RegCap	Region capture	Universal + Indexed Primer	AR019	GTGAAA (Added post-capture)

Step 4: Final PCR amplification and SPRI clean-up

Multiple PCR reactions were set up: x PCR reactions (x equals the remaining volume of Capture Hi-C library divided by factor 2.5) with 25µl each, as previously described in section 2.3.2, with 6 cycles of PCR. Following PCR all the reactions were pooled together in a fresh lo-bind tube. The beads were captured, the supernatant transferred to a fresh tube and the volume measured. The beads were retained and resuspended in 30µl 1xNEBuffer2 then stored in the freezer as a back-up.

SPRI beads (1.8 x volumes) were added to each sample and incubated at room temperature for 10-20 min. The beads were captured and washed twice on the magnet with fresh 70% ethanol. Beads were dried for 3 min at 37^{o}C. DNA was eluted in 100µl fresh TLE, the DNA transferred to a fresh lo-bind tube then the clean-up repeated by adding 180µl SPRI beads to the 100µl of library and incubated at room temperature for 10-20 min. The beads were captured and washed on the magnet twice with fresh 70% ethanol. Beads were dried for 3 min at 37^{o}C. DNA was eluted in 20-25µl TLE ensuring absolutely no beads were present in the sample.

Before sequencing, the quality and quantity of the libraries were checked by Bioanalyzer using a DNA-HS chip (section 2.2.2.2) and KAPA qPCR (section 2.2.2.3).

2.4. Sequencing of Capture Hi-C libraries

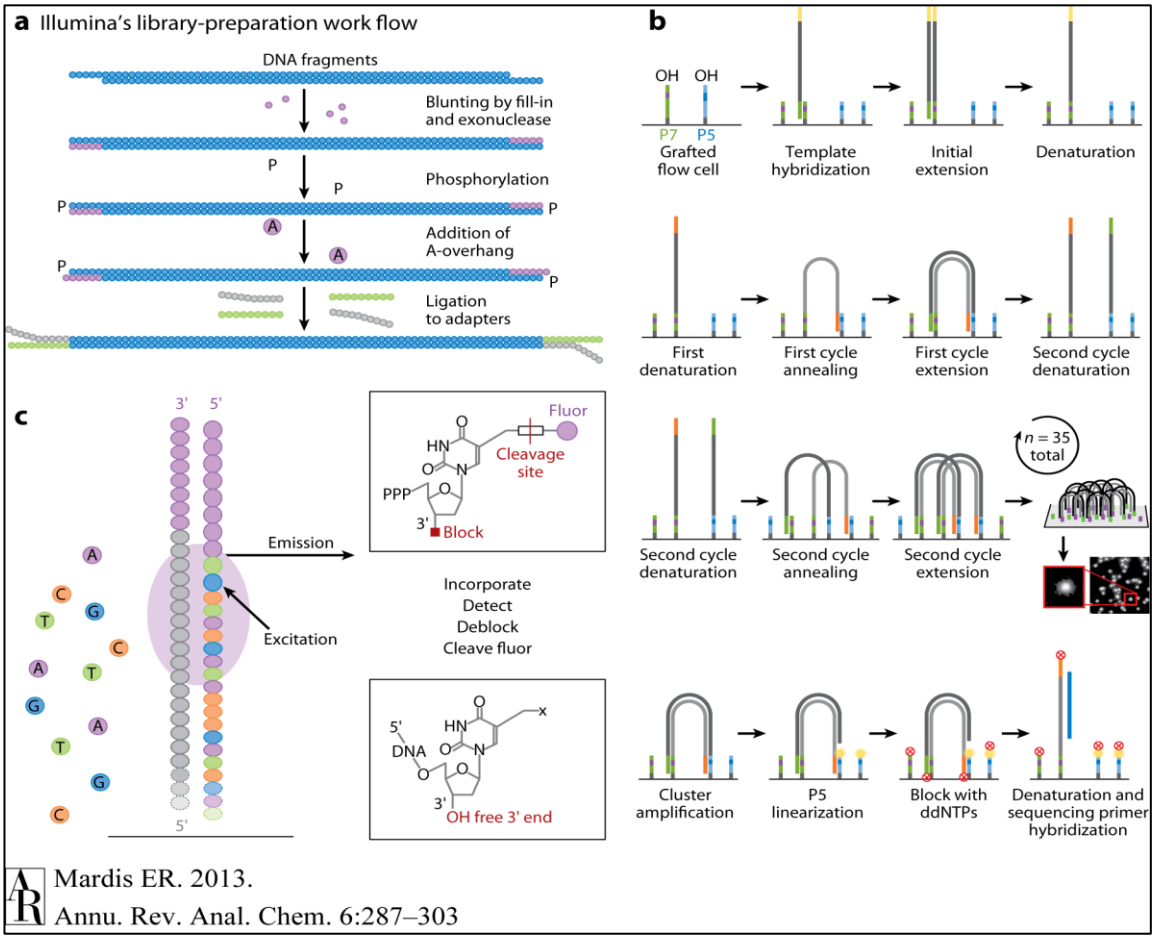
2.4.1. Principles of next-generation sequencing

The overall aim of DNA sequencing is to determine the order of nucleotides in a DNA molecule. The Human Genome Project (Lander *et al.* 2001; Sachidanandam *et al.* 2001) was the first major sequencing project using first generation Sanger sequencing (Sanger *et al.* 1977). This project took many years to complete, and since the completion in 2003, faster high-throughput methods have been developed known as next-generation sequencing (NGS) (Grada *et al.* 2013). NGS performs massively parallel sequencing of large stretches of DNA that has been fragmented into small sections allowing whole genomes to be sequenced in a single run. Adapters ligated to the ends of the library fragments are specific to the sequencing platform (Mardis 2013) and allow the amplification of the target DNA. High throughput instruments allow the multiplexing of reactions by utilising barcodes in the form of indexed adapters (discussed in sections 2.3.2 and 2.3.3) allowing the identification of each sample during data analysis.

2.4.2. Illumina technology

Illumina platforms use sequencing-by-synthesis to generate the nucleic acid sequence from the template library, as shown in Figure 13 (Mardis 2013; Quail *et al.* 2012). Diluted DNA libraries are immobilised onto a flow cell consisting of a glass slide that has adapters immobilised to the surface which are complementary to the adapter sequences used in the DNA libraries. Bridge amplification, using DNA polymerase, forms template DNA clusters on the flow cell. To determine the sequence of the template DNA, fluorescent reversible terminator bases (RT-bases) are added and unincorporated nucleotides are washed away. A camera takes pictures of the fluorescently labelled nucleotides then the dye and 3' terminal blocker are cleaved from the DNA allowing another cycle of amplification to begin. DNA chains are extended one nucleotide at a time allowing large numbers of clusters to be imaged with the same camera.

Figure 13: Illumina sequencing by synthesis

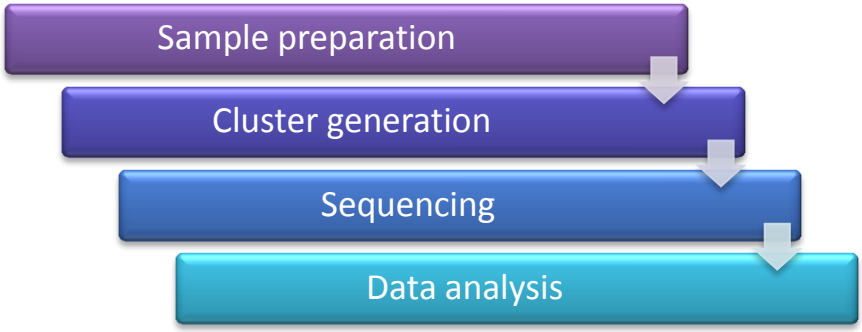


a) Illumina library construction, b) Cluster generation by bridge amplification, c) Sequencing by synthesis using reversible dye terminators

2.4.2.1. MiSeq and HiSeq platforms

The MiSeq is a bench-top sequencing instrument which has the advantage of being fast and simple to use for applications such as small genome and targeted sequencing. The flow cell for the MiSeq is a single lane so the amount of sequencing data obtained is a maximum of 15Gb (gigabytes), 25 million reads, and 2x300bp read length. The HiSeq is a large-scale, high-output instrument that can generate a maximum of 1000Gb of data, 4,000,000,000 reads, 2x125 read length in 2-11 days. The HiSeq flow cell consists of a flat glass slide containing 8 microfluidic channels which have adapter sequences immobilised to the surface that are complementary to the adapters used to make the libraries. Both platforms use a straightforward workflow, summarised in Figure 14.

Figure 14: Illumina sequencing workflow



2.4.3. Quality control using the Illumina MiSeq

Capture Hi-C libraries were quality-checked on a MiSeq using a V3 150 cycle kit (MS-102-3001, Illumina) prior to sequencing in greater depth on an Illumina HiSeq2500 using 50bp paired-end sequencing (Genomics Facility, Faculty of Life Sciences, University of Manchester). For the MiSeq runs, all reagent preparation and use of the instrument was carried out according to the Manufacturer's handbook.

Dilution and denaturation of DNA libraries

Dilution of DNA libraries was carried out according to the formula shown below, where *V(f)* is the final desired volume of the pool, *C(f)* is the desired final concentration of the DNA in the pool, # is the number of indexes and *C(i)* is the concentration of each sample going into the pool.

$$\text{Volume of Index} = \frac{V(f) \times C(f)}{\#\text{ x C(i)}}$$

An example calculation is shown below for a pool of 4 samples:

Component	V(f) µl	C(i) nM	C(f) nM	# Indexes	Vol (µl)
Sample 1	20	20	10	4	2.5
Sample 2	20	10	10	4	5
Sample 3	20	17	10	4	2.94
Sample 4	20	25	10	4	2
TLE buffer					7.56

For MiSeq V3 the final concentration of library was 4nM and for the HiSeq, the final concentration was 10nM. Samples for both instruments were prepared in a final volume of 20µl. The library preparation steps were carried out by myself for the MiSeq runs, the HiSeq samples were prepared for the HiSeq by the FLS sequencing facility.

For denaturation of the libraries 1ml of 0.2N sodium hydroxide (800µl Water + 200µl Stock 1.0N NaOH) was prepared in a 1.5ml tube and inverted several times to mix. To denature the DNA the components shown below were added to a 1.5ml tube, briefly vortexed, centrifuged at 280 x *g* for 1 min then incubated at room temperature for 5 min.

4nM sample DNA	5µl
0.2N NaOH	5µl

Following denaturation, 990µl pre-chilled HT1 was added to the tube containing denatured DNA to give a 20pM library which was placed on ice until ready to perform the final dilution. The denatured DNA was diluted to the desired concentration (15pM), inverted several times to mix, then briefly centrifuged. The diluted library was placed on ice until ready to load onto the reagent cartridge.

Final concentration	15pM
20pM denatured DNA	450µl
Pre-chilled HT1	150µl

Denature and dilute PhiX control

PhiX is used as an internal control. To dilute the PhiX to 4nM the following components were added to a 1.5ml tube:

10nM PhiX library	2µl
10mM Tris-HCl pH8.5, with 0.1% Tween-20	3µl

If not already prepared for denaturing samples within the last 12 hours, fresh 0.2N NaOH was prepared and 2nM PhiX library (5µl) and 0.2N NaOH (5µl) were combined in a fresh 1.5ml tube. The solution was briefly vortexed then centrifuged at 280 x *g* for 1 min. The mixture was incubated at room temperature for 5 min to denature the PhiX library.

Pre-chilled HT1 990µl was added to the tube containing denatured PhiX library to give a 20pM denatured library. To further dilute the PhiX library to 12.5pM, 375µl of 20pM PhiX library was added to 225µl Pre-chilled HT1.

Denatured PhiX control library (6µl) and denatured sample library (594µl) were combined in a 1.5ml tube (1% PhiX control, as recommended by Illumina) then placed on ice until ready to load onto the reagent cartridge.

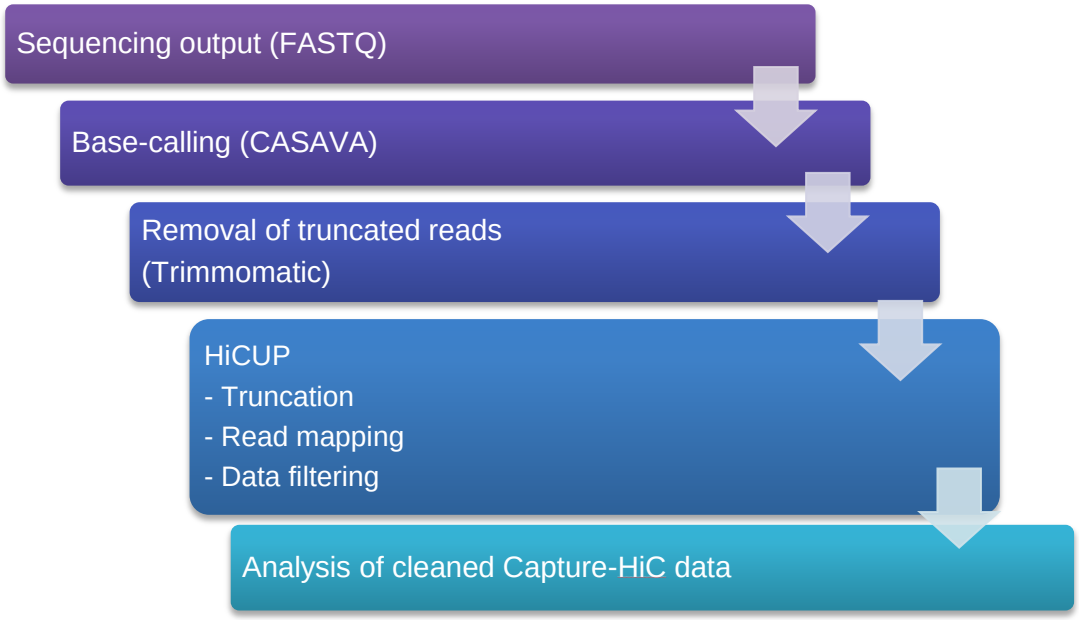
Loading sample libraries onto reagent cartridge

After preparing the reagent cartridge, the sample was loaded by piercing the foil seal over the reservoir labelled Load samples. Sample library (600µl) was loaded into the reservoir, without touching the foil. The MiSeq run was immediately set up using the MiSeq control software interface.

2.4.4. Sequencing QC

All sequencing QC, alignments and analysis (Sections 2.4.4.1 to 2.4.4.2) were carried out in-house by our bioinformatician, Mr Paul Martin. An outline of the analysis workflow is shown in Figure 15. Sequencing output from the MiSeq and HiSeq 2500 runs was in FASTQ format and CASAVA software (V1.8.2 Illumina) used to make base calls. Any reads that did not pass the Illumina filter were removed before further analysis. Trimmomatic (v0.30) (Bolger *et al.* 2014) was used to remove truncated or poor quality reads then the data was filtered using the Hi-C user pipeline HiCUP (Wingett *et al.* 2015). HiCUP is available for download at:
http://www.bioinformatics.babraham.ac.uk/projects/hicup/

Figure 15: Sequencing analysis workflow



2.4.4.1. HiCUP filtering

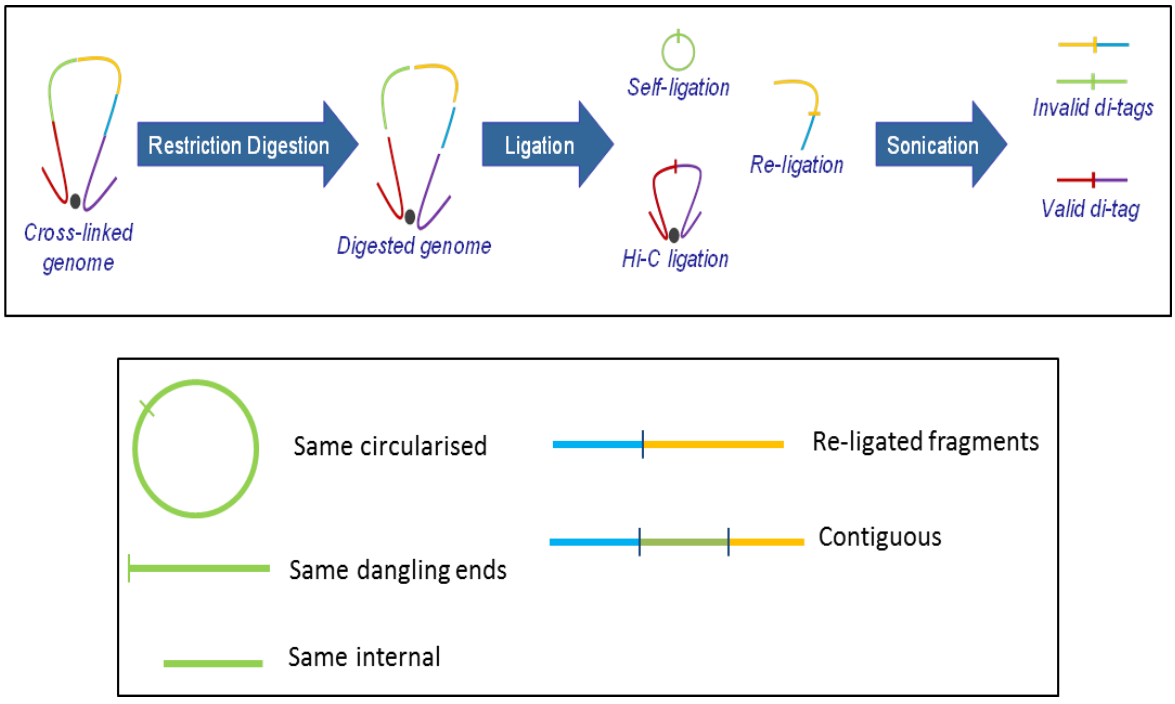
HiCUP maps paired-end sequencing data using the human GRCH37/hg19 reference genome sequence, using a set of PERL scripts which are run from the Linux/Unix command line. Parameters specific to Hi-C are used in the analysis which removes common experimental artefacts. Following HiCUP analysis, a QC report is generated which allows troubleshooting of various steps of the experiment.

Step 1: Truncating and mapping of Hi-C reads

Hi-C reads may contain sequences which map to more than one genomic region because ligation junctions can occur anywhere in the genome. Using the *Hin*dIII restriction fragments, 10bp ligation junctions were identified then the reads truncated at the putative Hi-C junction before alignment.

HiCUP uses BOWTIE (Langmead *et al.* 2009) to map Hi-C reads. The mapping is specific to Hi-C and uses very strict parameters to prevent mis-mapped reads. BOWTIE only allows reads mapping to one genomic region to be aligned. Forward and reverse Hi-C reads were mapped independently and then combined to create a di-tag containing parts of a different *Hin*dIII fragment on either end of the pair (Figure 16).

Figure 16: Hi-C ligation products



(Figure Adapted from Babraham Bioinformatics HiCUP presentation)

Step 2: Data Filtering

Data was filtered to remove invalid di-tags from the dataset and provided information about the overall library quality. Examples of invalid di-tags are fragments where both reads map to the same *Hin*dIII fragment making self-circularised products, or unligated products with dangling ends produced when fragments have been biotinylated but not ligated.

Any duplicate pairs, which are di-tags where both ends are present in another di-tag, were removed because they are experimental artefacts introduced during PCR amplification and could lead to false confidence in observed interactions. A high % of unique reads was desirable (>99%).

When the valid di-tags had been determined, the number of *cis* (close) reads was examined. Randomly ligated fragments containing no structural information produce libraries with a very high bias towards *trans* reads (~95%). A Hi-C library was considered very good quality if the % *Trans* was <50%.

2.4.4.2. Analysis of Capture Hi-C data

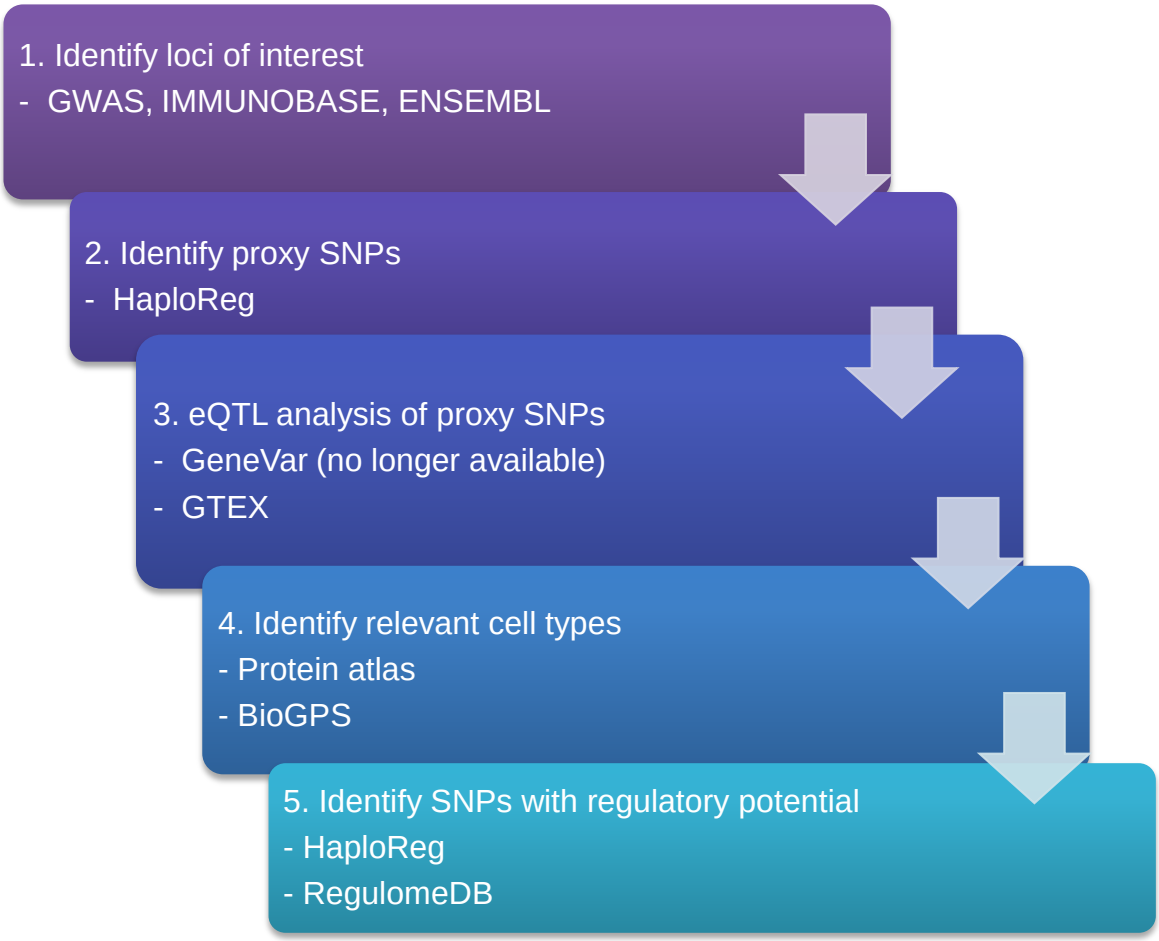
The analysis of the Capture Hi-C data is described in Martin *et al*. 2015. Following HiCUP filtering, off-target di-tags (where neither end mapped to a targeted *Hin*dIII restriction fragment) were removed using BEDTools (Quinlan *et al.* 2010) and command line tools. Di-tags separated by <20kb were also removed due to their very high interaction frequencies (Naumova *et al.* 2012). Di-tags were then assigned to four categories as described by Dryden *et al* (Dryden *et al.* 2014): (single baited *cis* <5Mb, single baited *cis* >5Mb, double baited *cis, trans*).

Significant interactions for *cis* within 5Mb were analysed using the 'High resolution analysis of cis interaction peaks' method (Dryden *et al.* 2014). Briefly, significant interactions were interactions above background seen in both biological replicates. Since more interactions occur by chance from regions of the chromosome that are close together, and more interactions are detected from baits that are 'more mappable', then the raw interaction reads were corrected for distance and probe 'mappability' based on the number of *trans* interactions observed (assumed to be random background noise). Corrected reads were then assessed for significance, based on the overall interaction frequency observed and using a false discovery rate (FDR) cut off of 5%. Interactions were considered statistically significant after combining replicates and filtering on FDR ≤5%. Statistically significant interactions were visualised using the WASHU Epigenome browser (Zhou *et al.* 2013b).

2.5. Bioinformatics

Publicly available bioinformatics databases were employed to identify and prioritise potential causal variants and genes identified in the Capture Hi-C experiments for further analysis as detailed in Figure 17.

Figure 17: Bioinformatics workflow, detailing bioinformatics tools employed for each step



Stage 1: Identify loci of interest

The 6q23 region associated with RA and other autoimmune diseases was selected for further intensive analysis. The RA associated region is intergenic, maps at some distance from the closest annotated gene, and is over 200kb from the best candidate gene (*TNFAIP3*), making it an ideal region to study long-range interactions in disease.

Stage 2: Identify proxy SNPs

The identification of an association signal in GWAS provides evidence that a particular genomic region could be implicated in disease. However, the most strongly associated SNP may not be the causal variant, which may be another SNP which is in strong linkage disequilibrium (LD) with it. The next stage of the bioinformatics workflow included searching genomic regions for highly correlated SNPs. There are three SNPS independently associated with RA in the 6q23 locus (Orozco *et al.* 2009), with my work focussing on an intergenic region, in a gene desert between *TNFAIP3* and *OLIG3*, containing the most strongly associated variant in the UK population, rs6920220.

HaploReg v4.1 (http://www.broadinstitute.org/mammals/haploreg/haploreg.php) (Ward *et al.* 2012; Ward *et al.* 2016) was used to search for proxy SNPs (r^2>0.8) associated with the rs6920220 variant in the *TNFAIP3-OLIG3* intergenic region. HaploReg uses LD information from the 1000 Genomes Project (Abecasis *et al.* 2012) to explore candidate regulatory SNPs at disease-associated loci. Details are provided about chromatin state (histone marks, DNase), protein binding annotation (Roadmap Epigenomics and ENCODE projects), the effect of SNPs on regulatory motifs, and the effect of SNPs on expression (eQTL studies).

Stage 3: eQTL analysis of proxy SNPs

The role of the proxy SNPs in regulating gene expression was investigated using eQTL databases which search publicly available eQTL data. The first database used was Genevar (Yang *et al.* 2010) based at the Sanger Institute (Cambridge, UK), which allows the visualisation of SNP-gene expression associations, however this database is no longer available. Genetic variation and gene expression profiles are available from a number of resources:

a) Adipose tissue, LCLs and skin from 856 healthy female twins, and from a subset of ~160 twins (Nica *et al.* 2011) from the MuTHER resource (Grundberg *et al.* 2012).

b) LCLs from 726 HapMap3 individuals from eight populations (Stranger *et al.* 2012).

c) Fibroblast, LCL and T-cells from the umbilical cords of 75 Geneva GenCord individuals (Dimas *et al.* 2009).

Gene-centric analysis was used to detect *cis*-eQTLs for *TNFAIP3* in HapMap CEU samples using a simple linear regression analysis. A SNP-centric analysis (*cis*-eQTL – SNP) was carried out on the rs6920220 SNP and its proxies in HapMap CEU samples, also using a simple linear regression analysis. SNP-gene associations were carried out for the rs6920220 SNP and proxies in HapMap individuals from CEU and other populations.

A second eQTL database used was the GTEx database (Genotype-tissue expression) (GTEx Project. 2013), which contains gene expression data from multiple human tissues.

Stage 4: Identify relevant cell types

Protein atlas (http://www.proteinatlas.org/) was used to examine the expression levels of a particular transcription factor or protein in different cell and tissue types; Biogps (http://biogps.org/) was used to provide information about a gene (such as expression in cell types, general information about the gene), for example, *TNFAIP3* expression levels.

Stage 5: Identify SNPs with regulatory potential

RegulomeDB (Boyle *et al.* 2012) and HaploReg v4.1 (Ward *et al.* 2016) were used to identify disease associated SNPs with regulatory potential within the 6q23 intergenic region. These databases annotate SNPs with regulatory elements in non-coding and intergenic regions using data from publicly available datasets (GEO, ENCODE, Epigenomics Roadmap and published literature), including variants mapping within markers of enhancers (histone modifications), transcription factor binding sites (ChIP-seq data) and open/active chromatin (DNaseHS), from international sources such as ENCODE, Epigenomics Roadmap and Blueprint. In RegulomeDB a score is given to each variant (1-6) based on functional evidence; variants with a low score are more likely to have a functional effect. The SNPs in LD with rs6920220 identified using HaploReg v4.1 were used in the RegulomeDB search.

2.6. Validation of long-range interactions by 3C-qPCR

Statistically significant interactions within the 6q23 locus were validated using 3C-qPCR.

2.6.1. Preparation of 3C libraries

For initial validation experiments, triplicate 3C libraries were prepared for GM12878 and Jurkat cell lines using the crosslinking, digestion, and ligation steps of the Hi-C protocol (detailed in section 2.3.2) excluding the biotin dATP fill-in. Further analysis was carried out using LCLs carrying the different rs6927172 alleles (G = risk, C = non-risk) (see Appendix Table 24 for cell lines used) and also with primary human synovial fibroblasts (provided by Dr. Caroline Ospelt, University Hospital of Zurich, Switzerland).

2.6.2. Primer design for 3C-qPCR

Primers for qPCR were designed in Primer 3 (Untergasser *et al.* 2012) for each set of interactions and checked in Primer-BLAST (Ye *et al.* 2012) for specificity (see Appendix Table 29). Short-range control interactions close to the anchor fragment, and primers in non-interacting regions were also designed. All primers were designed as close as possible to the restriction site in a unidirectional format (Dekker 2006). The short-range controls were used to control for digestion and ligation efficiency between samples and the negative control regions were included to show that the interaction in the target region was preferential.

2.6.3. Preparation of BAC Control Libraries

2.6.3.1. Growth of BAC cultures

BAC clones spanning the 6q23 locus (Figure 65 and Appendix Table 25) were supplied as a glycerol stock (Children's Hospital Oakland Research Institute; Life Technologies) which was streaked onto an LB agar plate containing 12.5µg/ml chloramphenicol. To make a starter culture, an isolated colony from a LB/chloramphenicol plate was picked using a sterile pipette tip and incubated in 5ml LB/chloramphenicol (12.5µg/ml) with shaking at 200rpm. After approximately 6 hours, 1ml of starter culture was added to 100ml LB/chloramphenicol (12.5µg/ml) broth and incubated with shaking at 200rpm overnight.

2.6.3.2. Isolation of BAC DNA using the Nucleobond BAC 100 kit

Bacterial cells were harvested by centrifuging in 50ml tubes at 1500 *x g* for 15 min at 4^{o}C and the supernatant discarded. To lyse the cells, the cell pellets were resuspended in 24ml Buffer S1 (containing RNaseA) making sure there were no clumps. The suspension was split into 2 x 12ml in 2 x 50ml tubes due to the large volumes required in subsequent steps. To both tubes 12ml Buffer S2 was added and the tubes mixed gently by inverting 6-8 times then incubated at room temperature for 2-3 min (maximum of 5 min). Pre-cooled Buffer S3 (12ml) was added to the suspension in both tubes then mixed gently by inverting 6-8 times until a homogeneous suspension containing an off-white flocculate formed then incubated on ice for 5 min. During the 5 min incubation, a Nucleobond BAC 100 column was equilibrated with 6ml Buffer N2, allowed to empty by gravity flow and the flow-through discarded. A Nucleobond filter was placed into a funnel and wetted with a few drops of buffer N2. The bacterial lysate was loaded onto the wetted filter, the cleared lysate loaded onto the equilibrated Nucleobond BAC 100 column, the column allowed to empty by gravity flow and the flow-through discarded. The column was washed with 2x18ml Buffer N3 and the flow-through discarded. The BAC DNA was eluted with 15ml Buffer N5 (preheated to 50^{o}C to increase yield) into a 50ml tube.

To precipitate the BAC DNA, 11ml room temperature isopropanol was added to the eluate. The samples were mixed carefully and transferred into 1.5ml tubes. The samples were centrifuged at 14,500 *x g* for 30 min at 4^{o}C. The supernatant was carefully removed, taking care not to dislodge the very small pellets. To wash the DNA 200µl room temperature 70% ethanol was added to each pellet and briefly vortexed. A few tubes were pooled together to fill a 1.5ml tube then centrifuged at 14,500 *x g* for 10 min at room temperature and the supernatant carefully removed. To wash the DNA a second time, 1ml room temperature 70% ethanol was added to one of the pellets and the same 1ml used to resuspend the pellets in the remaining tubes. The samples were then centrifuged at 14,500 *x g* for 10 min at room temperature. The ethanol was carefully removed with a pipette tip and allowed to dry at room temperature for 10-20 min. The DNA was resuspended in 400µl deionised water and dissolved using a shaking heat block at 37^{o}C for 10-60 min. The BAC DNA quantity and purity was analysed by Nanodrop and the identity of the BACs confirmed by PCR.

2.6.3.3. PCR confirmation of BAC clone identity

The suppliers of the BAC clones were not able to provide confirmation that the sequence of the BAC is correct, therefore the identity of each clone was checked by PCR using PCR primers designed to amplify two separate regions within the BAC.

PCR reactions were set up in a 25µl reaction using the recipe below. Bioline MyTaqHS polymerase and the supplied reaction buffer containing dNTPs were used for all reactions. PCR primers were diluted 1:10 from a 100µM stock solution and 0.5µl added to each reaction. As a

positive control, 100ng of human random control (HRC) DNA (Sigma) was amplified alongside the BACs. Water instead of DNA was used as a negative control. Following PCR, amplification products were electrophoresed at 120V for 75 min on a 1.5% agarose gel stained with ethidium bromide then visualised on a transilluminator. If any of the BAC clones did not amplify by PCR, but the positive control DNA did amplify, it was likely that it did not contain the correct sequence and an alternative BAC was ordered.

5x PCR reaction buffer	5µl
Primers	1µl (0.5µl each primer)
Water	17.5µl
Template/Water	1µl
Polymerase	0.5µl

PCR cycling was carried out in a thermocycler using the following parameters:

95°C	1 min	
95°C	15 secs	x35
55°C	15 secs	
72°C	10 secs	
72°C	10 min	
4°C	Hold	

2.6.3.4. BAC control library preparation

The protocol used to create the BAC control library used in the 3C-qPCR assays was based on the protocol by Naumova *et al* (Naumova *et al.* 2012).

a) Digestion of BAC genomic DNA

BAC DNA was digested in the following reaction overnight at 37°C with rotation. Equimolar amounts of each BAC clone were used in the digest.

BAC DNA (equimolar amounts of each BAC clone)	20µg
10× CutSmart buffer	50µl
HindIII (or up to 10% of total reaction volume)	50µl
Water	(up to final volume of 500µl)

After digestion, DNA was purified using a 1:1 phenol:chloroform:isoamyl alcohol extraction in 2ml PLG light tubes as previously described. The upper phase was transferred to a new 2ml tube and 1/10 volume of 3M sodium acetate, pH5.2 was added and the tube vortexed briefly. Ice-cold 100% ethanol (2.5 x volumes) was added and the tube gently inverted to mix. Samples were incubated at −20°C for a few hours to precipitate then centrifuged at full speed in a microfuge at 4°C for 20 min. The pellet was washed twice in 1ml of 70% ethanol, the supernatant removed and the pellet briefly air-dried. The pellet was resuspended in 44µl water and incubated at 37°C for 15 min to dissolve the DNA.

b) BAC DNA ligation

The ligation reaction was prepared as follows and incubated at 16°C overnight in a thermoblock.

Digested BAC DNA	44µl
10× T4 DNA ligase buffer	6µl
T4 DNA ligase	5µl
Water	5µl
Total volume	60µl

After overnight ligation, the samples were incubated at 65°C for 15 min to inactivate the ligase.

c) Purification of BAC genomic DNA control template

Water (140µl) was added to the ligation reaction to make the final volume 200µl then the DNA purified by 2 x 1:1 phenol:chloroform:isoamyl alcohol extractions and 1 x 1:1 chloroform extraction in 2ml PLG Light tubes and the DNA precipitated as previously described. The pellet was washed twice in 1 ml of 70% ethanol, the supernatant removed and the pellet briefly air-dried. The pellet was resuspended in 100µl TE buffer and incubated at 37°C for 15 min to dissolve the DNA. The 3C control template was stored at −20°C.

2.6.4. 3C-qPCR

All qPCR was performed in triplicate in 384-well optical plates using Power SYBR green (Life Technologies) on a QuantStudio 12K Flex instrument (Life Technologies). 50ng of 3C library template was used per reaction and a no-template (water) control (NTC) was included.

Each assay included the following primer sets:

- anchor primer + test primers
- anchor primer + negative control regions (NCR)
- anchor primer + short range (SR) controls for normalisation

For each set of interactions a standard curve was generated using 10-fold serial dilutions from 50ng of the BAC control library generated in section 2.6.3.4. The standard curve was used to generate Intercept and Slope values which are used when calculating relative interaction frequencies to adjust for differences in primer efficiency. The standard curve is also a positive assay control because the BAC library should generate all possible interactions within the region being investigated.

The following recipe and cycling parameters were used in all assays.

Template	1µl
Primers (diluted to 10µM)	0.5µl each
2x SYBR green mastermix	5µl
Water	3µl
Total reaction volume	10µl

50°C	2 min	
95°C	10 min	
95°C	15 sec	⎫
60°C	1 min	⎭ 40 cycles

The relative interaction frequency (RIF) was calculated as described by Hagege *et al* (Hagege *et al.* 2007) using the calculations detailed below.

1. Standard curve calculated to obtain Slope and Intercept values

2. Average C_T values determined for all interactions

3. Intercept value (b) subtracted from Average C_T value

4. Value from step 3 divided by slope value (a)

5. Calculated $10^{(C_T\text{-Intercept/a})}$

6. Data was normalised by dividing the test region value from step 5 by the short-range interaction value to give the RIF

A positive result was obtained if the RIF in the test region had a statistically significant higher RIF than the non-interacting NCR.

Example	b Intercept	a Slope	Average C_T	C_T - Intercept	C_T - Intercept/a	10^ C_T - Intercept/a	Normalised
SR_2	25.27	-3.1912	27.942	2.6716663	-0.837	0.1454	
SNPs_1	23.696	-3.699	28.890	5.1936662	-1.404	0.0394	0.271097
NCR_2	24.729	-3.6711	32.483	7.7535001	-2.112	0.0077	0.053108

2.7. Investigation of regulatory protein binding in the 6q23 region by ChIP

2.7.1. Introduction to ChIP

Chromatin immunoprecipitation (ChIP) is a popular method used to study DNA-protein interactions *in vivo* to determine whether specific elements, such as transcription factors, are associated with a certain genomic sequence. The various ChIP assays, summarised in Table 4, have been comprehensively reviewed (Christova 2013; Collas 2010; Fullwood *et al.* 2009b) and some of the main applications of ChIP are summarised below (Figure 18). The protocol is outlined in the schematic (Figure 19).

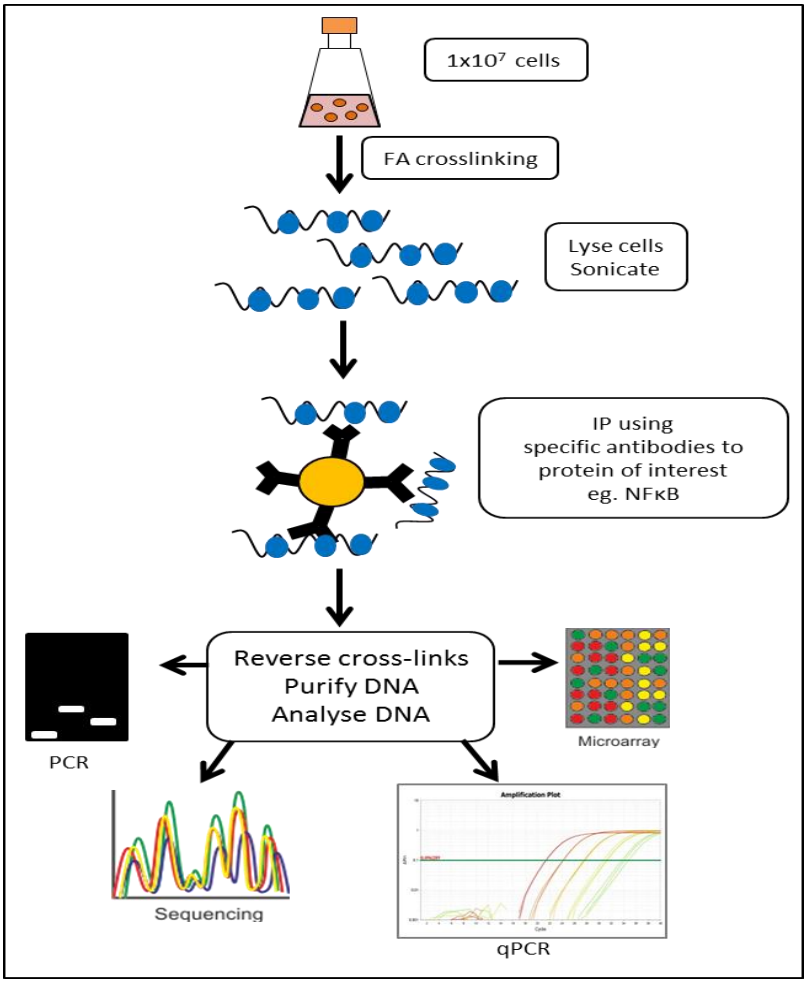
Figure 18: Applications of the ChIP assay

Before designing a ChIP assay it is important to consider if the target protein is expressed in the cell line of choice, how abundant the protein is, the binding affinity of the target protein to DNA (Table 7) and if suitable antibodies are available for the target protein.

Table 7: Cell numbers and applications of ChIP

Abundance of protein target	Molecules per locus	Cells required per ChIP	Examples
High	High	10^4	Modified histones, RNApolII
Medium	Medium	10^5-10^6	General TFs
Low	Low	10^6-10^7	Sequence specific TFs
Low	Indirect binding	10^7 or more	Accessory factors

Figure 19: Schematic of the ChIP assay



For each cell line, optimisation of various steps should be carried out (Table 8)**.**

Table 8: Optimisation of the ChIP assay

ChIP Step	Parameters to Optimise	Reason
Crosslinking	Time, FA concentration	Over/under-crosslinking could result in missing interactions
Shearing	Method (enzyme/mechanical), duration, power settings	Obtaining chromatin of the correct size-range, type of ChIP assay (some prefer enzymatic methods)
Magnetic beads	Protein A or G, mixture of A/G, volume	For optimum antibody affinity and reduction of background noise
Antibody	Concentration	Ratio of pulldown over background – too high Ab = more background
IP duration	1-6 hours room temperature or 4^{o}C overnight	Efficient pulldown of high-abundance epitopes = shorter IP time

2.7.2. Selection of transcription factors and antibodies for ChIP assays

It was predicted through bioinformatics analysis that the transcription factors NF-κB and BCL3 (B-cell lymphoma 3-encoded protein) bound to rs6927172, the SNP with the most evidence of regulatory potential in the *TNFAIP3-OLIG3* region in 6q23. ChIP assays for BCL3 and the NF-κB subunits p50 and p65 were performed to detect genotype specific differences in NF-κB binding in cell lines containing the three different genotypes of rs6927172. ChIP assays were also carried out using antibodies to H3K4me1 and H3K27ac to confirm if the target SNPs lied in an enhancer region.

Antibody selection is crucial for successful ChIP. Many manufacturers (such as Abcam and Millipore) have developed validated ChIP-grade antibodies which have proven performance in ChIP. ChIP-grade antibodies were available from Abcam for the transcription factors NF-κB p50 and p65, and for the histone marks of active enhancers H3K4me1 and H3K27ac**.** A ChIP-grade antibody was available from Santa Cruz for the transcription factor BCL3.

2.7.3. Detailed ChIP protocol

Step1: Crosslinking of DNA-Protein interactions

Cells were grown to ~80-90% confluence, counted using a CASY automated cell counter, and $1x10^7$ cells prepared for crosslinking. The cell suspension was transferred to a 50ml conical-bottomed tube and centrifuged at 200 *x g* for 5 min at room temperature and the medium discarded. The pellet was resuspended in 5ml pre-warmed PBS, centrifuged at 200 *x g* for 5 min at room temperature, and the PBS discarded. Cells were resuspended in 40ml PBS then formaldehyde (37% stock solution) added to a final concentration of 1% and the cells incubated on a rocking platform for 10 min at room temperature, then placed on ice immediately. The formaldehyde was quenched by the addition of 2M ice-cold glycine to a final concentration of 0.125M, and the cells incubated on a rocking platform on ice for 5 min. The cells were pelleted by centrifuging at 450 *x g* for 5 min at 4^{o}C and the supernatant discarded. At this point cells were transferred to cryovials and lysed.

Step 2: Cell lysis and chromatin shearing

Cells were lysed in 1ml of complete ChIP lysis buffer (50mM Tris-HCl pH8.1, 10mM EDTA, 1% SDS) containing fresh protease inhibitors for 15-30 min on ice then the lysed cells transferred to Nunc cryovials, snap-frozen in liquid nitrogen and stored at -80°C.

For chromatin shearing, a Covaris S220 was used which uses Adaptive Focused Acoustic (AFA) technology to mechanically shear the DNA. The following variables were used:

Target BasePairs	200-400bp
Duty Cycle	5% for LCL; 10% for Jurkats
Peak incident power	140 Watts
Cycles per burst	200
Temperature	4^{o}C
Time	20-25 min (10µl aliquot taken for QC check every 5 min)

To verify that the chromatin fragments were within the correct size-range, 90µl cold TE and 2µl 10mg/ml RNaseA was added to each 10µl sample and incubated for 30 min at 37^{o}C with shaking at 400rpm on a thermomixer. Proteinase K (5µl of 20mg/ml stock) was added to each sample and incubated for 2 hours at 65^{o}C with shaking at 400rpm. DNA was purified using a QIAquick PCR purification kit according to the manufacturer's protocol and eluted in 25µl EB. Samples were visualised on a 1.5% agarose gel stained with ethidium bromide.

To remove any cell debris, sheared chromatin was centrifuged in a microfuge for 15 min at maximum speed at 4^{o}C and the supernatant removed to a fresh tube. The supernatant was the total chromatin used for the ChIP experiment. Cleared chromatin was aliquoted into 100µl samples (100µl per IP containing 1x10^{6} cell equivalents of lysate) and stored at -80^{o}C for up to 3 months.

Step 3: Immunoprecipitation (IP)

Each IP was carried out on triplicate chromatin aliquots to provide technical replicates. Negative control was a no-antibody control or IP with non-specific antibody (anti-GFP). Non-IP'd chromatin (Input) was used for qPCR normalisation to generate a % Input value. Biological replicates were performed on Jurkat T-cells and rs6927172 genotype-specific HapMap individuals: CC (n=10), GG (n=3), CG (n=8). Lo-bind tubes and pipette tips were used throughout to prevent the loss of material.

For each sample an excess of Protein G, Protein A or a mixture of A and G Dynabeads® was prepared (Table 9). Beads were washed three times with 50mg/ml BSA/PBS then incubated for 15-30 min on the last wash on a rotator to block any non-specific binding to the beads. After the final wash, the beads were captured on the magnet and the wash buffer removed then the beads resuspended in the starting volume of PBS/BSA. An excess of dilution buffer (16.7mM Tris HCl pH8.1, 1.1% Triton X-100, 0.01% SDS, 167mM NaCl) containing protease inhibitors was prepared (400µl of complete dilution buffer used per IP). The chromatin aliquots were thawed on ice then dilution buffer containing protease inhibitors added to each sample, making a total volume of 500µl. At this point 5µl (1%) of diluted chromatin was removed as 1% INPUT control and stored at 4^{o}C. Washed magnetic beads were added to each sample, specific antibody added (Table 9) then the samples incubated overnight at 4^{o}C with rotation.

Table 9: Parameters used in ChIP assays

Antibody	Beads	Volume of Beads	Conc. Ab
NFκB-p50 (ab7971)	Protein A	20µl	8µg
NFκB-p65 (ab7970)	Protein A	20µl	8µg
BCL-3 (sc-185)	Protein A/G mix	20µl	8µg
H3K4me1 (ab8895)	Protein A/G mix	20µl	1µg
H3K27ac (ab4729)	Protein A/G mix	20µl	1µg

Step 4: Elution and reverse cross-linking

Following overnight incubation the beads containing the immobilised DNA-Antibody complex were captured on the magnet and the supernatant removed. The complex was washed by resuspending the beads in 0.5ml each of the cold buffers listed in the order below. Each wash was carried out for 3-5 min on a rotator followed by magnetic clearance and removal of supernatant.

a) Low salt wash buffer (0.1% SDS, 1.0% Triton X-100, 2mM EDTA, 20mM Tris-HCl (pH 8.1), 150mM NaCl)

b) High salt wash buffer (0.1% SDS, 1.0% Triton X-100, 2mM EDTA, 20mM Tris-HCl (pH 8.1), 500mM NaCl)

c) LiCl wash buffer (1.0% Igepal-CA630 (NP-40), 1.0% sodium deoxycholate, 1mM EDTA, 10mM Tris-HCl (pH 8.1), 250mM LiCl)

d) TE buffer (10mM Tris-HCl (pH 8.1), 1mM EDTA)

The beads were resuspended in 100µl elution buffer (1% SDS, 0.1M $NaHCO_3$) and 1µl proteinase K (20mg/ml stock) added to reverse crosslinks in all the samples including the input samples. Samples were incubated for 2 hours with shaking at 62^{o}C, followed by 95^{o}C for 10 min then cooled to room temperature. The beads were captured on the magnet and the supernatant removed to a fresh tube. DNA was purified using a QIAquick PCR purification kit according to the manufacturer's protocol, then all samples eluted in 50µl water and stored at -20^{o}C.

Step 5: qPCR analysis of ChIP enrichment

To detect the relative enrichment of regions interacting with the target protein, qPCR of ChIP and Input samples was carried out. All qPCR was carried out in triplicate in 384-well optical plates using SYBR green on an Applied Biosystems QuantStudio 12K Flex qPCR instrument. Primers were designed for the target SNP region, positive and negative control regions and the efficiency of each primer pair validated using human genomic DNA (see Appendix Table 30 for primer sequences). The efficiency was calculated from the slope of the standard curve using the following equation:

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

A reaction that was 100% efficient should have a slope of -3.32, however efficiencies between 90-110%, corresponding to a slope of -3.58 to -3.10 were acceptable. Melt-curves were carried out for each assay to ensure primer specificity. A no-template control (NTC) was amplified alongside each primer set, in each assay. The following recipe and cycling parameters were used for SYBR green assays.

Template	1µl
Primers (diluted to 10µM)	0.5µl each
2x SYBR green mastermix	5µl
Water	3µl
Total reaction volume	10µl

50°C	2 min	
95°C	10 min	
95°C	15 sec	⎫
60°C	1 min	⎭ 40 cycles

Step 6: Analysis of qPCR data

Following qPCR, the % Input for each sample was calculated in Excel following the example shown below in Table 10 (from Life Technologies).

Statistical analysis using T-Tests was carried out to determine statistically significant differences in regulatory protein binding to the different SNP genotypes. P values <0.05 were considered statistically significant.

Table 10: Example of ChIP qPCR analysis to obtain percentage input values

		Step 1			**Step 2**
		***Adjusted input to 100%**			**Percent input**
	Raw C_T	(C_T Input - 6.644)		Triplicate average C_T	100*2^(Adjusted input - C_T (IP)
Input (1%)	32.7	26.1	Adjusted input	26.1	
			No Ab control	34.6	0.3
			Sample #1	31.3	2.7
			Sample #2	29.9	7.2

* For example, if the starting input fraction was 1%, then a dilution factor (DF) of 100 or 6.644 cycles (i.e., log2 of 100) was subtracted from the C_T value of the diluted input.

3. Results Section 1

Analysis of long-range interactions by Capture Hi-C

3.1. SNP selection for Capture Hi-C

In order to design custom baits for the Capture Hi-C the target regions were defined for the four autoimmune diseases under investigation. All RA, JIA and PsA loci which reached genome-wide significance were included in the design. The associated regions contained the lead SNP and all SNPs in LD (r^2>0.8) with the lead SNP. For the RA loci, the regions include SNPs identified by LD and Credible sets analysis with the overlapping regions merged. Details for the RA loci are shown in Table 11, which shows the size of the LD region for each locus, the index SNP and the number of SNPs in LD. Details for the JIA and PsA loci are included in Appendix Table 35. Also included in the Appendix is a combined list of all the RA, PsA and JIA loci put forward into the region capture (Table 36). The T1D credible SNPs were defined by a collaborator (Dr Chris Wallace, NIHR Cambridge Biomedical Research Centre, Cambridge Institute for Medical Research, University of Cambridge) and a list provided to include in the capture.

Table 11: RA SNP selection for Capture Hi-C

Locus	Start	End	Interval size	Index SNP	No. of SNPs in LD
1p36.32	2483960	2721149	237189	rs2843401	145
1p36.1	17673101	17674402	1301	rs2240336	3
1p34.3	38614866	38644861	29995	rs883220	9
1p13.2	114303807	114377568	73761	rs2476601	2
1p13.1	117280695	117280696	1	rs798000	1
1q23.3	161463875	161483977	20102	rs10494360	27
1q32.1	198779680	198810008	30328	rs2014863	13
2p16.1	61077102	61170913	93811	rs34695944	20
2p14	65556323	65598906	42583	rs6546146	18
2q11.2	100636756	100744683	107927	rs10209110	27
2q32.3	191900448	191935804	35356	rs13426947	9
2q33.2	204604602	204777818	173216	rs11571302	71
2q33.2	204604602	204777818	173216	rs1980422	70
3p14.3	58183635	58318477	134842	rs35677470	4
4p15.2	26085479	26128710	43231	rs932036	26
4q27	123030582	123289204	258622	rs78560100	112
5q11.1	55436850	55442249	5399	rs71624119	5
5q21.1	102595836	102681586	85750	rs39984	52
6q21	106435343	106508640	73297	rs6911690	60
6q23	137959234	138006504	47270	rs6920220	9
6q25.3	159489790	159496713	6923	rs629326	3
6q27	167537753	167537754	1	rs59466457	1
7q32.1	128575551	128581835	6284	rs3807306	8
8p23.1	11337586	11353110	15524	rs4840565	18
9p13.3	34707372	34755359	47987	rs2812378	5
9q33.2	123640499	123721510	81011	rs10739580	84
10p15.1	6098948	6108340	9392	rs10795791	6

10p15.1	6390449	6404700	14251	rs947474	14
10p14	8095339	8097368	2029	rs2275806	2
10q21.2	63781257	63813790	32533	rs12764378	8
11p12	36480985	36529349	48364	rs570676	15
11q12.2	60888000	60934812	46812	rs595158	40
11q23.3	118718728	118746433	27705	rs4938573	28
12q13.3-14.1	58012110	58105094	92984	rs10683701	128
15q14	38828139	38847763	19624	rs8043085	14
15q23	69984461	70010647	26186	rs8026898	6
16q24.1	86004871	86021624	16753	rs13330176	13
17q12	37912376	38080912	168536	rs12936409	90
19p13.2	10427720	10492274	64554	rs34536443	3
20q13.2	44730244	44749251	19007	rs6032662	12
21q22.12	35909624	35930915	21291	rs2834512	13
21q22.12	36695908	36745167	49259	rs9979383	6
22q12.3	37544244	37545505	1261	rs3218251	5

chr7	27090828	27303174	HOXA - Control region
chr16	104265	263351	HBA - Control region

3.2. Generation of Capture Hi-C libraries

To study the role of long-range chromatin interactions in RA associated regions, Hi-C libraries were generated for use in Capture Hi-C experiments.

3.2.1. Hi-C library quality control

Biological replicate Hi-C libraries from the GM12878 (LCL) and Jurkat T-cell lines were prepared according to the protocol described in Section 2.3.2. At the same time, 3C control libraries were generated to act as a comparison during Hi-C library PCR digest analysis. Quality control (QC) of the Hi-C libraries was carried out at several points to ensure that the libraries were of sufficient quality to progress through to sequencing.

a) Library quantification

Hi-C and 3C control libraries were quantified by Qubit using the Quant-IT dsDNA BR kit. All libraries produced a high quantity of DNA (a range of 625-875ng/ul) which was used in subsequent steps of the Hi-C protocol (Appendix Table 37).

b) Visual inspection of library integrity by agarose gel electrophoresis

Aliquots of Hi-C and 3C library were analysed by gel electrophoresis to check the integrity of the libraries. Figure 20 shows representative gels confirming that Hi-C and 3C libraries from both the GM12878 (Figure 20A) and Jurkat (Figure 20B) cell lines produced tight bands above 10kb with no degradation products.

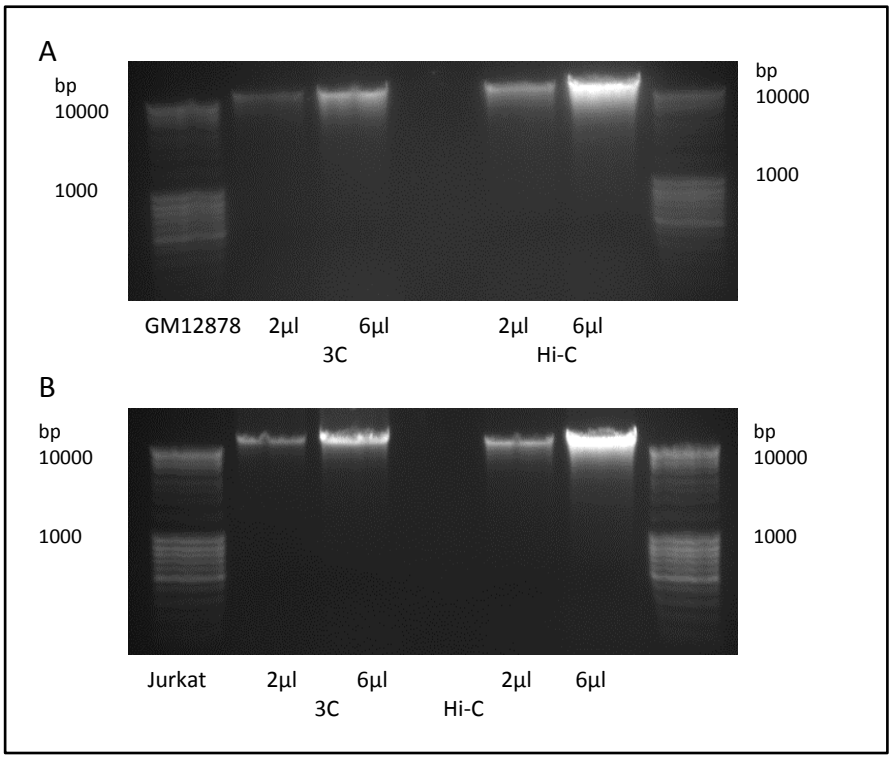
c) Detection of short-range and long-range interactions by PCR

PCRs to detect previously described short-range and long-range interactions were carried out on all Hi-C and 3C libraries from the GM12878 (Figure 21A-B) and Jurkat (Figure 21C-D) cell lines using a range of primers. Short range interactions using published *Hin*dIII or Dekker AHF control primers (Belton *et al.* 2012; Lieberman-Aiden *et al.* 2009) were observed in all samples. Short range interactions within the RA SNP region were also confirmed. Long-range interactions using primers at increasing distance from the *Myc* promoter were also demonstrated in all the samples tested (primer sequences obtained from Dr Stefan Schoenfelder, Babraham Institute, Cambridge, UK).

d) PCR digest assays to assess biotinylation and fill-in reaction efficiency

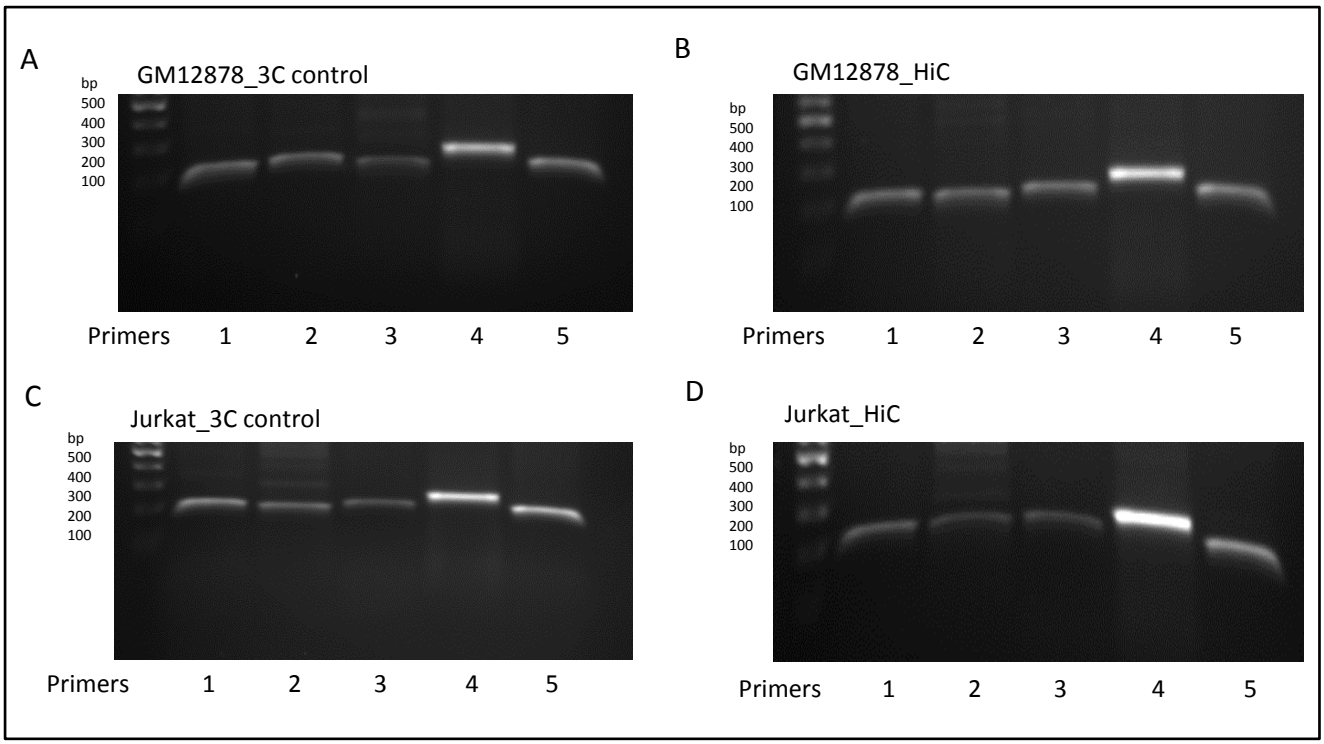
The successful biotinylation and fill-in during the Hi-C experiment produces a new restriction site for *Nhe*I. PCR products from amplifications using short range control primers were digested with *Hin*dIII, *Nhe*I and both *Hin*dIII and *Nhe*I along with an undigested control. The 3C libraries for both Jurkat (Figure 22A and C) and GM12878 (Figure 22B and D) cell lines showed digestion in only the reactions containing *Hin*dIII, as expected. The Hi-C libraries (Jurkat Figure 22A and C; GM12878 Figure 22B and D) showed digestion in only the reactions containing *Nhe*I, confirming that the biotinylation and ligation reactions had been successful.

Figure 20: Agarose gel analysis of 3C and Hi-C libraries



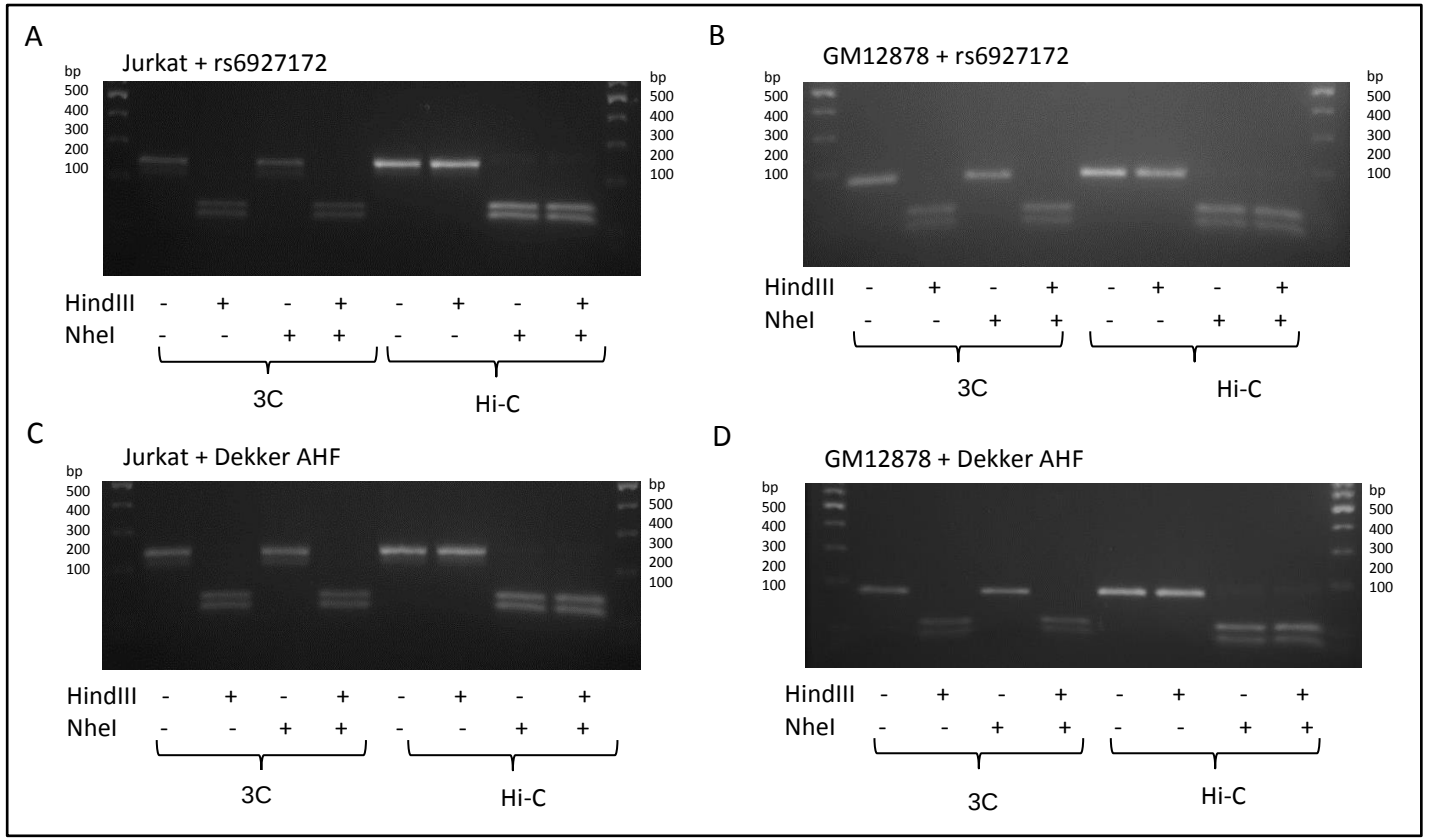
The first step of Hi-C library QC is to visualise aliquots from a 1:10 dilution of both Hi-C and 3C control libraries on a 0.8% agarose gel stained with ethidium bromide. Libraries should run as a tight band above 10kb in size. Low molecular weight (MW) smearing is indicative of a poor quality, degraded library which would be unsuitable for sequencing. MW marker = MassRuler DNA ladder (Life Technologies). The results shown for GM12878 (A) and Jurkat (B) are representative of the results obtained from this QC step.

Figure 21: PCRs to detect short-range and long-range interactions



Short-range and long-range interactions were detected using PCR in 3C and Hi-C libraries (see Appendix Table 25). Each PCR reaction amplified 200ng of template DNA. PCR products were visualised on a 1.5% agarose gel stained with ethidium bromide. The results shown in the above gels are representative results from library preparations (A) GM12878_3C, (B) GM12878_HiC, (C) Jurkat_3C, (D) Jurkat_HiC. Primer pairs used are (1) rs6927172_1_F + rs6927172_2_R, (2) AHF64 + AHF66, (3) Human_Myc_G2 + Human_Roger_1R, (4) Human_Myc_G2 + Human_Myc_O3, (5) Human_Myc_G2 + Human_Myc_540. Molecular weight marker = MassRuler DNA ladder.

Figure 22: PCR digest assays of GM12878 3C and Hi-C samples



PCR digests using short-range primers were used to amplify 3C and Hi-C libraries in order to assess the efficiency of the biotinylation and fill-in reactions. All PCRs used 200ng of template DNA per reaction. PCR products were visualised on a 1.5% agarose gel stained with ethidium bromide. The results shown in the above gels are representative results from library preparations (A) Jurkat amplified with rs6927172 primers, (B) GM12878 amplified with rs6927172 primers (C) Jurkat amplified with Dekker AHF primers, (D) GM12878 amplified with Dekker AHF primers. Molecular weight marker = MassRuler DNA ladder. Below each gel indicates which restriction enzymes were used in the digest and which are 3C or Hi-C samples. 3C libraries only digested with *Hin*dIII and Hi-C libraries only digested with *Nhe*I as expected.

3.2.2. Biotin pulldown and pre-capture quality control

a) Removal of biotin from non-ligated ends

To prevent the isolation of non-ligated biotinylated products from the Hi-C library it is necessary to carry out a biotin removal step. The protocol specifies to carry out biotin removal on aliquots of Hi-C library up to a maximum of 40µg (eight aliquots of 5µg). For each library, the maximum number of aliquots was processed (See Appendix Table 37).

b) Post-size selection quantification

Following size selection of sheared and end-repaired Hi-C libraries, the quantity of library was determined by Qubit using the Quant-IT dsDNA BR kit. To avoid overloading the streptavidin beads used in the biotin pulldown it was important to not use more than 2.5µg of DNA. If the quantity of DNA exceeded 2.5µg, multiple pulldowns were carried out as detailed in Appendix Table 38 and the samples pooled at the end.

c) Test amplifications

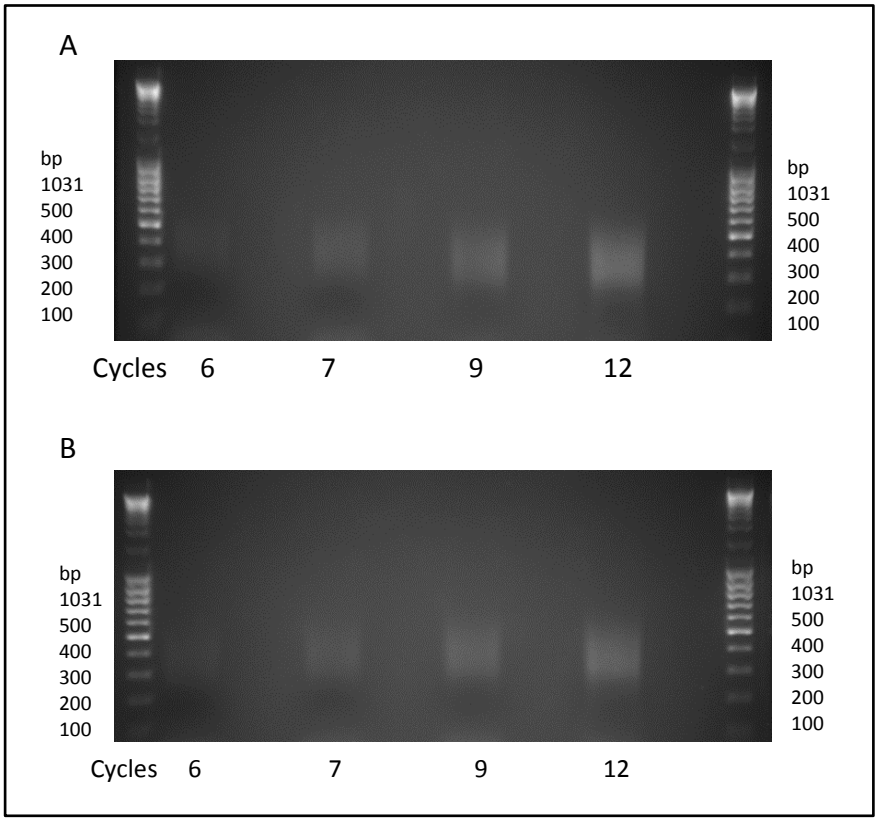
To determine the correct number of PCR cycles for final library amplification, test PCRs were performed using 6, 7, 9 and 12 cycles. Representative gels showing the results of test amplifications for GM12878 libraries and Jurkat libraries are shown in Figure 23. For both libraries a very faint product was detected at 6 cycles, gradually increasing in intensity as the number of cycles was increased, producing a very strong product at 12 cycles. For final library amplification, 8 cycles of PCR was carried out.

d) Bioanalyzer assessment

The quality and quantity of the Hi-C libraries was checked using a DNA-HS Bioanalyzer chip. Figure 24 and Figure 25 show the electropherograms and Bioanalyzer gels for Hi-C libraries prepared from the GM12878 and Jurkat cell lines. Details of the average library size and library concentration are shown in Table 12. The GM12878 libraries had an average size of ~386bp and Jurkat libraries ~416bp (optimal size ~400bp). The biological replicate GM12878 libraries had an average size of ~390bp and Jurkat libraries ~396bp.

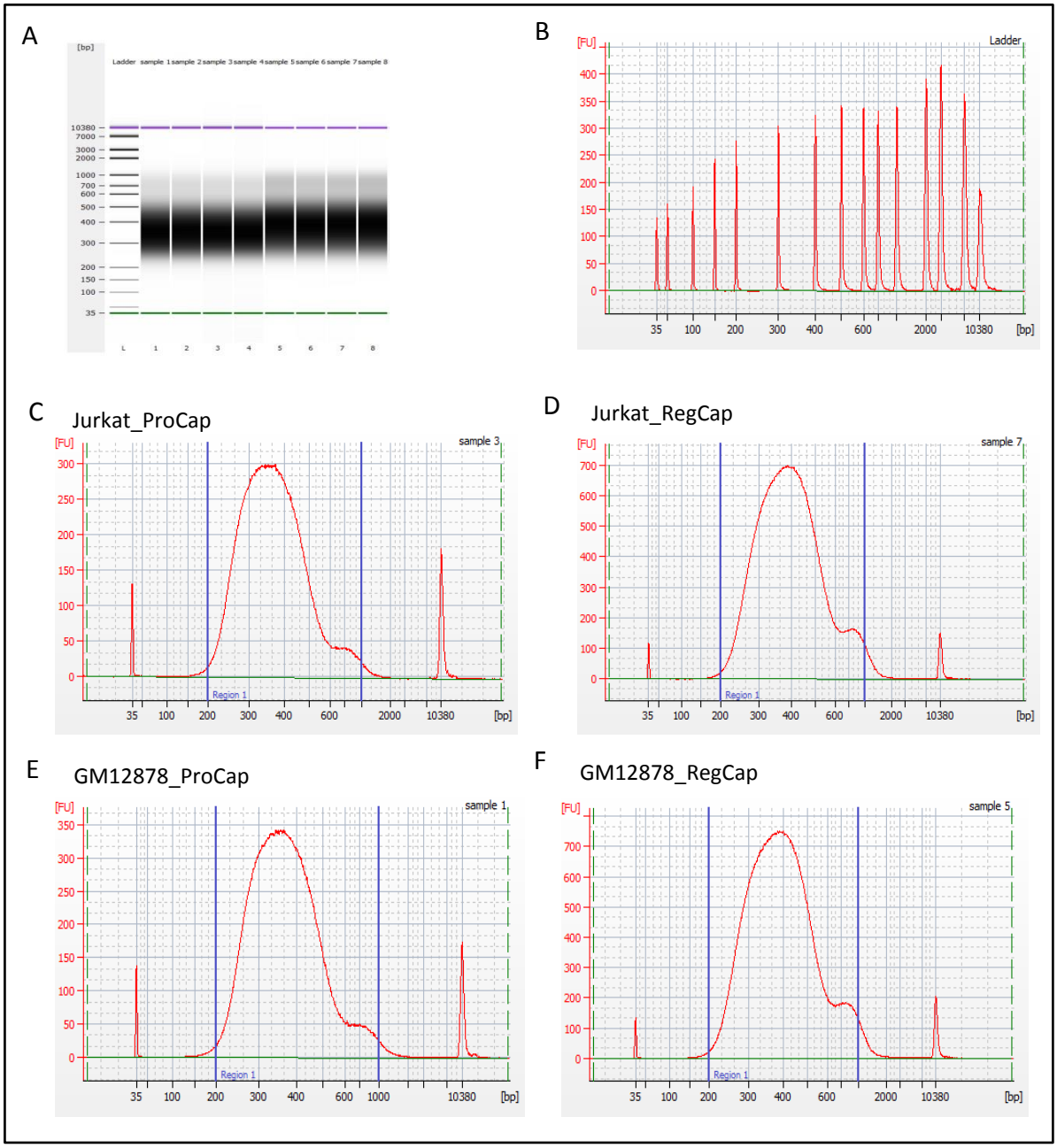
All the QC steps showed that the libraries were of excellent quality and the samples were carried through to the capture hybridisation steps. The first replicates provided 750ng of material per capture, however, the yield from the biological replicate libraries was not as high as the first samples so less material was available for the captures.

Figure 23: Test amplifications of Hi-C libraries



Hi-C libraries were amplified from DNA immobilised on streptavidin beads using pre-capture TruPE_PCR_1.0.33 and TruPE_PCR_2.0.33 primers (Appendix Table 27). PCR was performed using 6, 7, 9 and 12 cycles using Phusion polymerase (NEB) and the products visualised on a 1.5% agarose gel stained with ethidium bromide. The gels shown are representative of the results obtained from test amplifications of GM12878 (A) and Jurkat (B) Hi-C libraries. Faint products were detected at 6 cycles, and final amplification was carried out using 8 cycles to prevent over-amplification. Molecular weight marker = MassRuler DNA ladder.

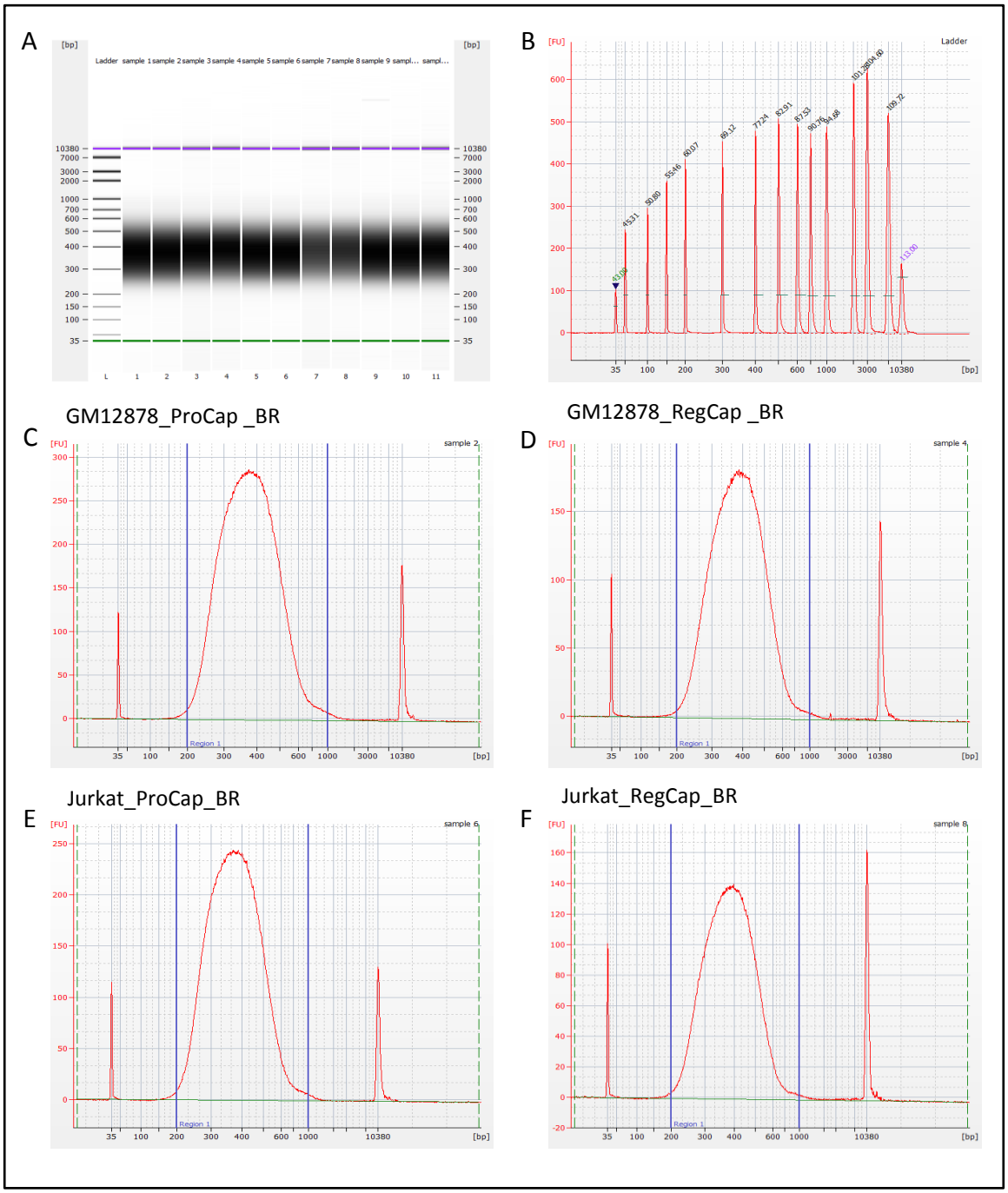
Figure 24: Bioanalyzer assessment of pre-capture Hi-C libraries (first biological replicate)



Duplicate samples of Hi-C library diluted 1:5 were loaded onto an Agilent Bioanalyzer DNA-HS chip to determine the quantity and average sizes of the libraries prior to solution hybridisation.

(A) Bioanalyzer gel showing all the samples loaded onto the chip in duplicate, (B) Electropherogram of molecular weight marker. C-F show the Bioanalyzer electropherograms for each sample (C) Jurkat Promoter Capture Hi-C library, (D) Jurkat Region Capture Hi-C library, (E) GM12878 Promoter Capture Hi-C library, (D) GM12878 Region Capture Hi-C library.

Figure 25: Bioanalyzer assessment of pre-capture Hi-C libraries (second biological replicate)



Duplicate samples of Hi-C library diluted 1:5 were loaded onto an Agilent Bioanalyzer DNA-HS chip to determine the quantity and average sizes of the libraries prior to solution hybridisation.

(A) Bioanalyzer gel showing all the samples loaded onto the chip in duplicate, (B) Electropherogram of molecular weight marker. C-F show the Bioanalyzer electropherograms for each sample (C) GM12878 Promoter Capture Hi-C library, (D) GM12878 Region Capture Hi-C library, (E) Jurkat Promoter Capture Hi-C library, (D) Jurkat Region Capture Hi-C library.

Table 12: Bioanalyzer results for GM12878 and Jurkat Hi-C libraries

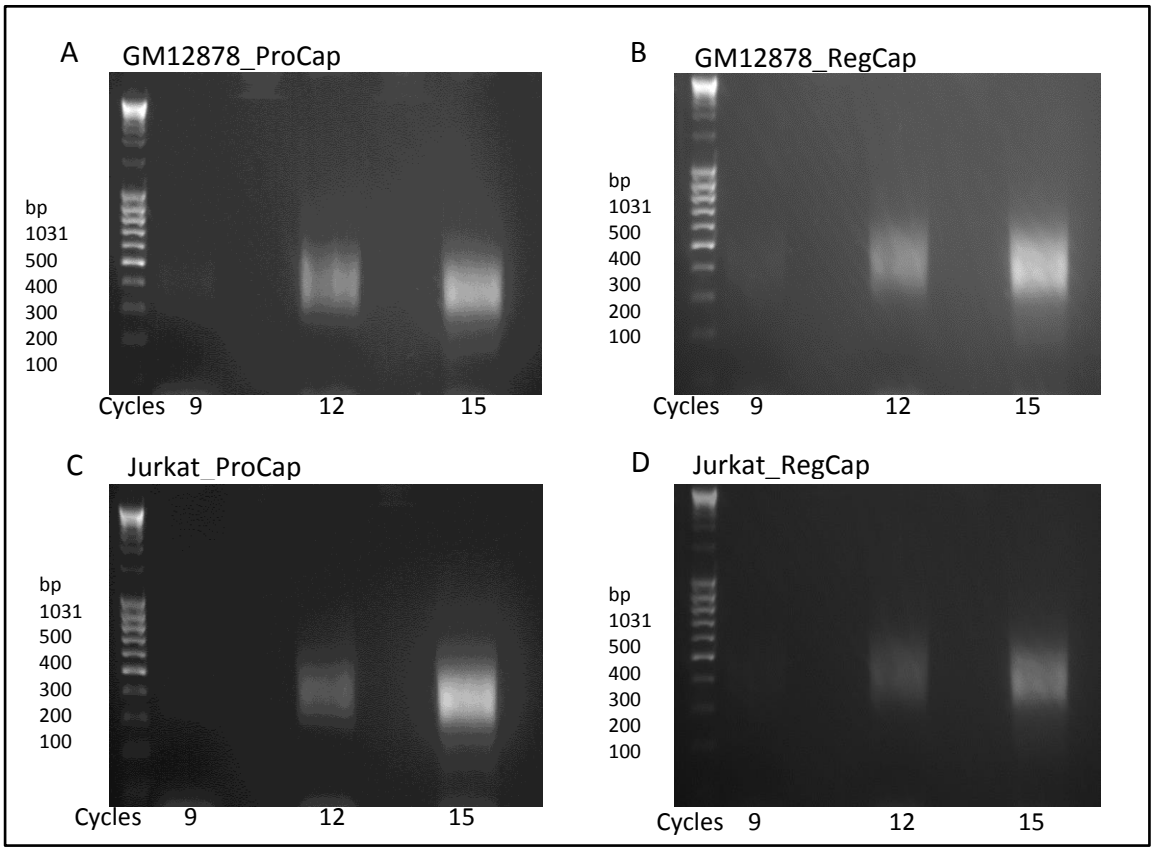
Sample	Average Size (bp)	Molarity (pmol/l)	Conc (ng/µl)	Conc x5 (ng/µl)	Total yield in 38µl	For 750ng (µl)	Average for 750ng (µl)
GM12878_ProCap_A	386	29,695.2	6.8	34.0	1,292	22.00	21.20
GM12878_ProCap_B	387	31,997.4	7.3	36.7	1,395	20.40	
GM12878_RegCap_A	385	25,456.7	5.8	29.2	1,108	25.00	25.00
GM12878_RegCap_B	386	26,085.4	6.0	29.9	1,137	25.00	
Jurkat_ProCap_A	418	49,758.80	12.2	60.8	2,309	12.30	11.45
Jurkat_ProCap_B	413	58,057.60	14.1	70.4	2,673	10.60	
Jurkat_RegCap_A	417	61,577.20	15.1	75.3	2,860	9.90	10.25
Jurkat_regCap_B	420	57,402.30	14.1	70.4	2,673	10.60	
GM12878_ProCap_BR	391	16,586.60	4.30	21.5	731	**23.3**	**500ng**
GM12878_RegCap_BR	388	22,489.20	4.80	24	816	**20.8**	**500ng**
Jurkat_ProCap_BR	396	9,090.70	2.60	13	442	**30.8**	**400ng**
Jurkat_RegCap_BR	396	10,437.60	2.50	12.5	425	**32**	**400ng**

Data for the average size, concentration and molarity of the Hi-C libraries from the Bioanalyzer assessment are shown, along with the concentration adjusted for the dilution factor, total yield and the amount of sample required for the capture experiments. Each sample was tested in duplicate (biological replicate samples – only one replicate tested) and the average used to calculate the amount of library to put into the capture. The capture experiments ideally required 750ng of material but the biological replicate samples had lower yield so less input could be used for the captures.

3.2.3. Solution hybridisation – Capture Hi-C

Following library QC, Hi-C libraries for GM12878 and Jurkat cell lines were used in Promoter Capture and Region Capture experiments (Sections 2.3.1 to 2.3.3). Post-Capture libraries were quality-assessed prior to next-generation sequencing. Test amplifications using either 9, 12 or 15 cycles were carried out to determine the optimum number of cycles needed to generate the final library for Illumina sequencing. For all libraries a faint smear at 9 cycles was observed (Figure 26), so 6 cycles was used for the final amplification of each captured Hi-C library. To ensure both captures were carried out from the same sample, the final amplification PCRs were split equally to create two separate samples – one for each capture. The number of samples used in the final amplification reactions, volumes recovered and the samples created are listed in Appendix Table 39.

Figure 26: Test amplifications of post-capture libraries (first biological replicates)



Hi-C libraries were amplified from DNA immobilised on streptavidin beads using post-capture Universal and barcoded primers (Appendix Table 27). To determine the correct number of cycles for final amplification of post-capture libraries, test amplifications using 9, 12 and 15 cycles were carried out and visualised on 1.5% agarose gels stained with ethidium bromide. Representative gels from the first Capture experiments are shown in the results above. Faint products at 9 cycles were observed for all samples: (A) GM12878 Promoter Capture library, (B) GM12878 Region Capture library, (C) Jurkat Promoter Capture library, (D) Jurkat Region Capture library. Molecular weight marker = MassRuler DNA ladder.

3.2.4. Post-Capture library quality control

To obtain an accurate size and quantity of the libraries, post-capture QC was carried out by both Bioanalyzer and KAPA qPCR.

a) Bioanalyzer

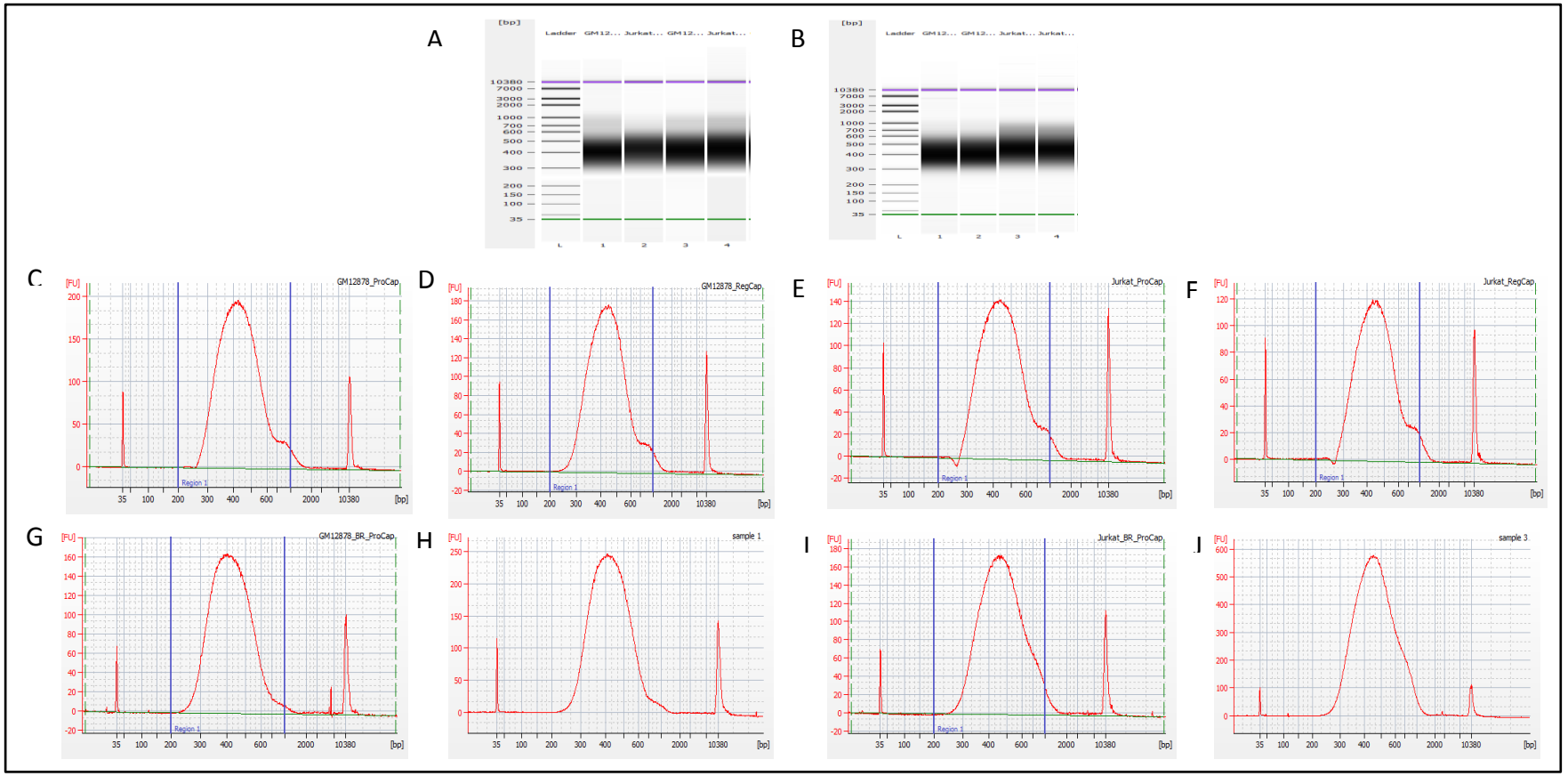
Results for the Promoter and Region capture libraries are shown in Figure 27. The Bioanalyzer results confirmed that the libraries were of the correct size range for Illumina sequencing (<500kb) and of good quality. The library size was larger than in the pre-capture libraries because of the addition of extra bases introduced by the final PCR amplification to enable sequencing of the libraries.

b) KAPA qPCR

KAPA qPCR was used to obtain an accurate quantification of the libraries. Samples were diluted 1:1000 then serially diluted 2-fold up to 1:8000. Representative standard curves and amplification plots for the qPCR assays are shown in Figure 28 and the calculated concentrations shown in Table 13 and Table 14. Quantification of the GM12878 libraries by qPCR showed concentrations that were very close to the Bioanalyzer results, indicated by efficiency values close to 1.0. The Jurkat Promoter and Region capture libraries also gave close results to the Bioanalyzer. The concentration obtained from the most concentrated sample to fit on the standard curve was used in the final calculations.

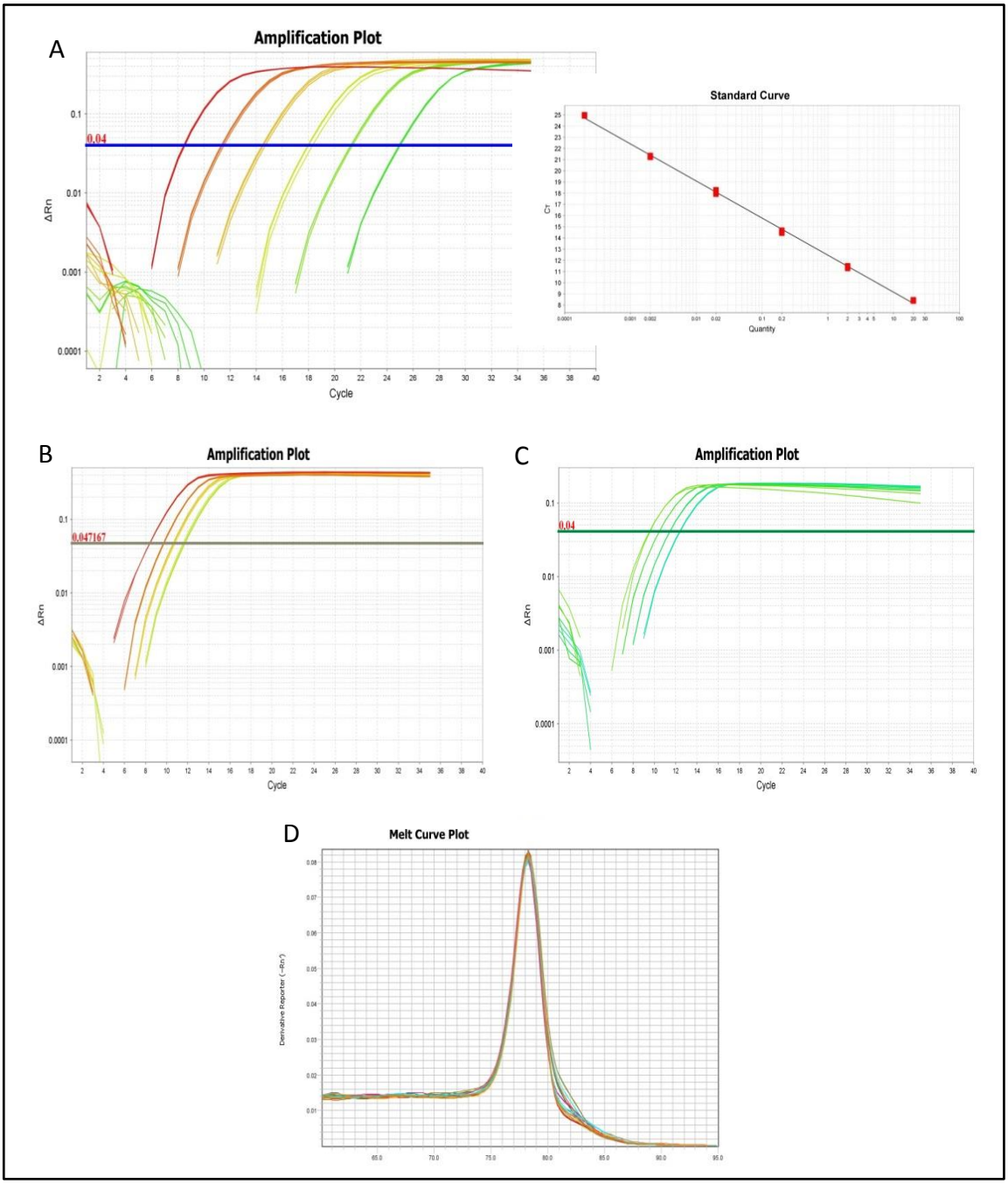
The sizes and quantification obtained from the Bioanalyzer and KAPA qPCR were used to normalise the concentrations of the libraries to ensure accurate dilution and pooling for Illumina sequencing, shown in Appendix Tables 40-42. The final dilutions needed to make a sample of 10nM in a volume of 20µl were calculated and these samples were carried forward to sequencing on the HiSeq 2500.

Figure 27: Post-capture Bioanalyzer assessment of Capture Hi-C libraries



Samples of undiluted captured Hi-C library were loaded onto a DNA-HS chip to determine the quantity and average sizes of the libraries following capture experiments. (A) Bioanalyzer gel of first Capture Hi-C libraries, (B) Bioanalyzer gel of biological replicate Capture Hi-C libraries, C-J show the Bioanalyzer electropherograms for each library (C) GM12878 Promoter Capture, (D) GM12878 Region Capture, (E) Jurkat Promoter Capture, (F) Jurkat Region Capture, (G) GM12878_BR Promoter Capture, (H) GM12878_BR Region Capture, (I) Jurkat_BR Promoter Capture, (J) Jurkat_BR Region Capture.

Figure 28: Kapa qPCR analysis of Post-Capture libraries



Capture Hi-C libraries were quantified using dilutions ranging from 1:1000 to 1:8000. The plots shown in the figure are representative results. (A) Example amplification plot and standard curve generated from the KAPA qPCR standards, (B) Representative amplification plots for GM12878 libraries (C) Representative amplification plots for Jurkat libraries, (D) Melt curve analysis was also performed to ensure the specificity of the qPCR primers.

Table 13: Post-Capture Library quantification (first samples)

	Dilution	Rep_1	Rep_2	Rep_3	Average	Average Fragment Length (bp)	Size adjusted conc (pM)	[KAPA qPCR] (nM)	[Bioanalyzer] (nM)	Efficiency
GM_ProCap	1000	23.399	22.003	21.786	22.396	452	22396	22.40	15.13	1.48
	2000	9.067	8.786	9.06	8.971	452	17942	17.94	15.13	1.19
	4000	4.768	4.617	4.593	4.659	452	18637	18.64	15.13	1.23
	8000	2.397	2.258	2.275	2.31	452	18480	18.48	15.13	1.22
GM_RegCap	1000	19.674	19.898	19.589	19.720	459	19419	19.42	12.03	1.61
	2000	7.960	7.778	8.192	7.976	459	15710	15.71	12.03	1.31
	4000	4.297	3.827	4.052	4.058	459	15987	15.99	12.03	1.33
	8000	2.210	2.052	1.935	2.065	459	16273	16.27	12.03	1.35
JK_ProCap	1000	12.404	12.211	13.142	12.585	466	12207	12.21	9.69	1.26
	2000	6.327	6.295	6.371	6.331	466	12281	12.28	9.69	1.27
	4000	3.675	3.749	3.256	3.56	466	13812	13.81	9.69	1.43
	8000	1.797	1.512	1.829	1.712	466	13289	13.29	9.69	1.37
JK_RegCap	1000	12.349	12.238	11.675	12.087	470	11624	11.62	9.367	1.24
	2000	5.483	5.888	6.025	5.798	470	11153	11.15	9.367	1.19
	4000	3.154	2.927	3.061	3.047	470	11722	11.72	9.367	1.25
	8000	1.546	1.787	1.802	1.711	470	13168	13.17	9.367	1.40

Table 14: Post-Capture Library quantification (Biological replicate samples)

	Dilution	Rep_1	Rep_2	Rep_3	Average	Average Fragment Length (bp)	Size adjusted conc (pM)	[KAPA qPCR] (nM)	[Bioanalyzer] (nM)	Efficiency
GM_BR_ProCap	1000	21.011	21.478	21.766	21.418	438	22102	22.10	12.20	1.81
	2000	8.515	8.684	8.629	8.609	438	17769	17.77	12.20	1.46
	4000	3.888	3.905	3.926	3.906	438	16124	16.12	12.20	1.32
JK_BR_ProCap	1000	13.996	13.888		13.942	491	12834	12.83	9.90	1.30
	2000	6.581	6.524	6.535	6.547	491	12053	12.05	9.90	1.22
	4000	3.375	3.291	3.299	3.322	491	12231	12.23	9.90	1.24
GM_BR_RegCap	1000	15.847	16.267	16.509	16.208	437	16763	16.76	12.60	1.33
	2000	6.39	6.639	6.611	6.547	437	13542	13.54	12.60	1.07
	4000	3.126	3.258	3.244	3.209	437	13277	13.28	12.60	1.05
JK_BR_RegCap	1000	50.846	50.7	51.04	50.862	481	47795	47.80	32.10	1.49
	2000	18.96	19.46	20.7	19.707	481	37037	39.41	32.10	1.23
	4000	17.179	17.29	17.39	17.286	481	64976	69.15	32.10	2.15

3.3. Sequencing QC

Sequencing data was mapped and filtered by our in-house bioinformatician using HiCUP to remove any experimental artefacts generated by the Hi-C experiment. HiCUP generated Quality Control reports of the data which could be used to assess the quality of both the sequencing and the libraries. The reports were used to construct summary tables of the data and plots to illustrate the data in excel. Analysis of the libraries on the MiSeq showed that the quality was suitable for full sequencing on the HiSeq, based on the high percentage of valid and unique di-tags and the low percentage of background *trans* interactions (MiSeq sequencing summary data included in Appendix 1 Tables 43-45 and summarised in Figure 66).

3.3.1. Truncation and mapping

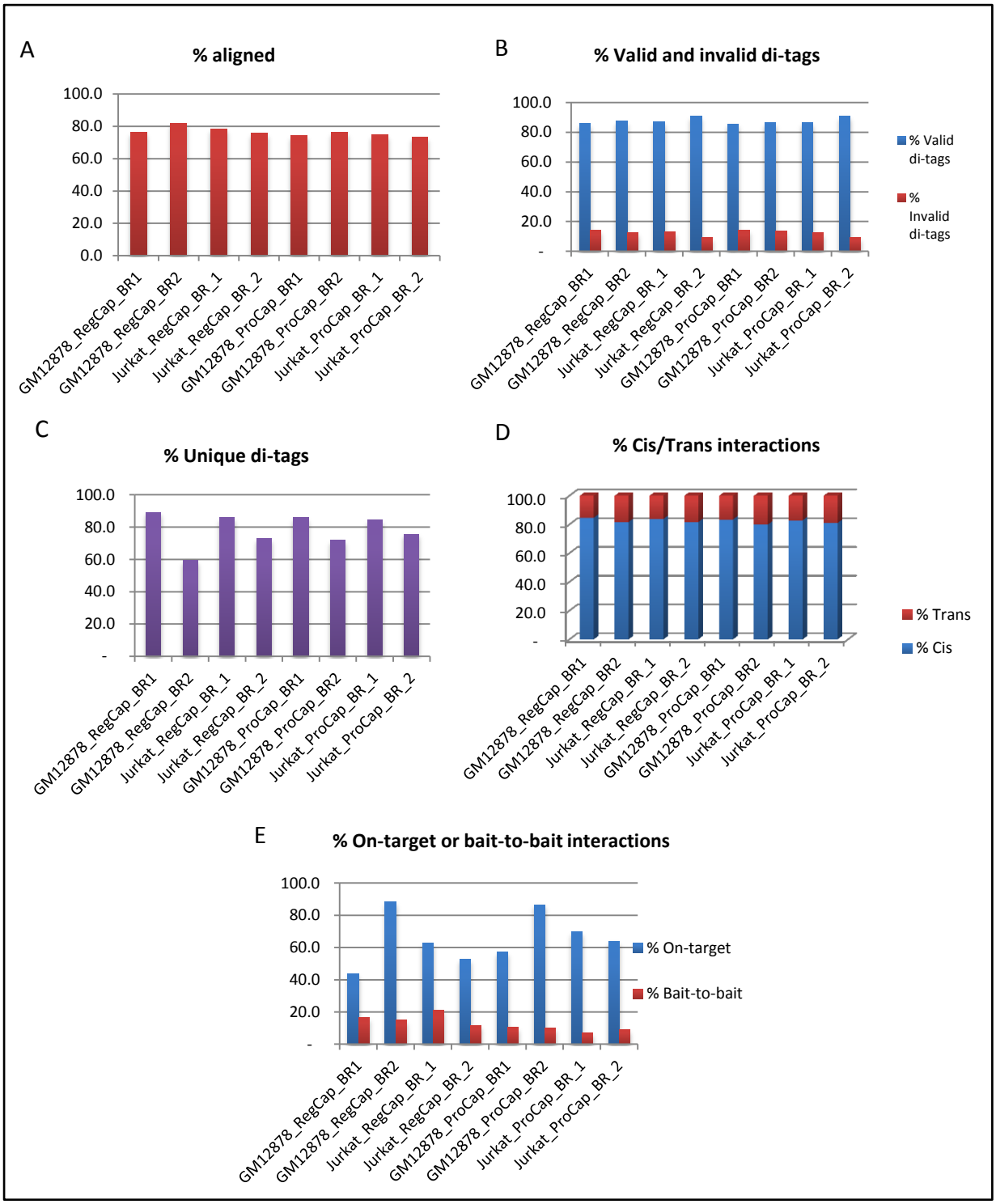
Processing of the sequencing data started with truncation of the sequencing reads at putative Hi-C ligation junctions allowing the reads to be mapped to the reference genome. The sequence mapping data showed how many reads were processed and the number of reads that generated a di-tag containing one *Hin*dIII fragment from a capture target region and one *Hin*dIII fragment from its ligated interaction partner.

Promoter Capture libraries were sequenced on one lane of a HiSeq 2500 and Region Capture libraries sequenced on 0.5 lane of a HiSeq 2500. The sequencing QC statistics generated from HiCUP for the HiSeq runs are shown in Appendix 1 (Tables 47-50). The total number of reads processed for the Region Capture libraries ranged from 76 million to 108 million with an average of 91 million reads. The total number of reads processed for the Promoter Capture libraries ranged from 148 million to 181 million with an average of 164 million reads. From the total reads processed, approximately 76% of the reads could be aligned to the reference genome (Figure 29A).

3.3.2. Filtering

Filtering of data was carried out to remove invalid di-tags from the datasets and provide information about overall library quality. From the total number of aligned di-tags, the number of valid and invalid di-tags was determined. Of the aligned di-tags processed, the percentage of valid di-tags containing parts of a different *Hin*dIII fragment on either end of the pair across all the samples was an average of 87.6% (Figure 29B). The percentage of unique valid di-tags was an average of 78.2% (Figure 29C). The percentage of *cis* interactions averaged across all the samples was 82.4%. The percentage *trans* was an average of 17.6% (Figure 29D).

Figure 29: HiSeq quality data from HiCUP reports

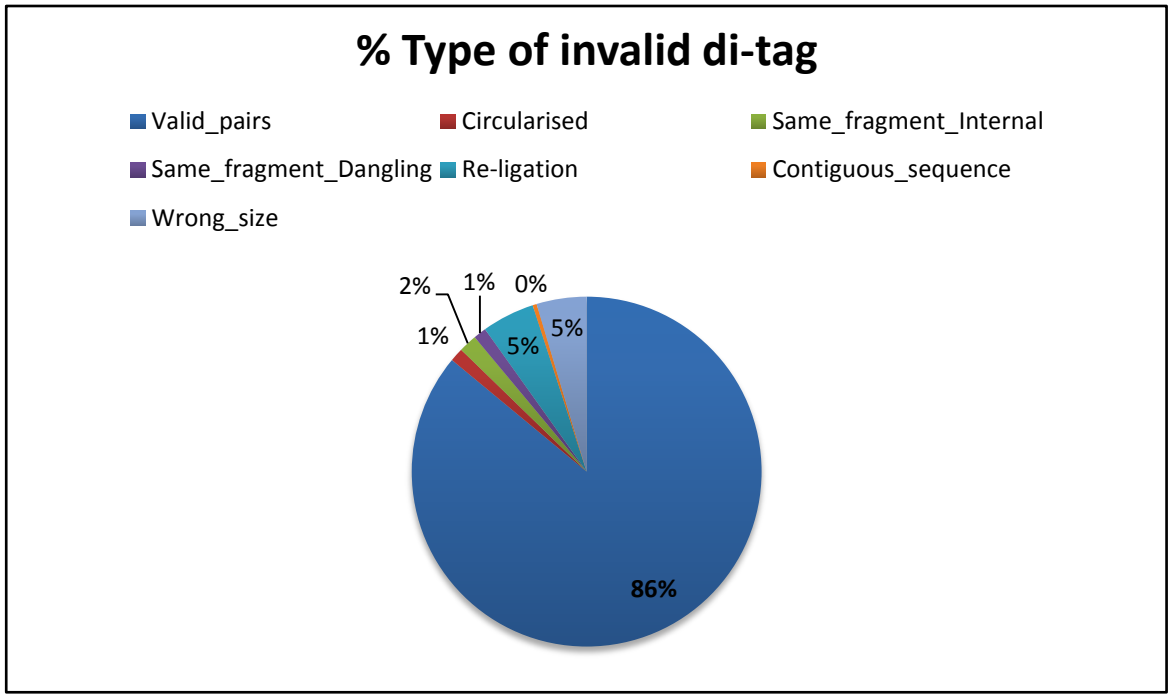


The HiCUP reports generated from the HiSeq sequencing data was used to produce plots of the summary statistics. The percentage of aligned reads from all the sequencing runs is shown in (A), the percentage of valid/invalid di-tags is shown in (B), the percentage of unique di-tags is shown in (C), the ratio of *cis/trans* is shown in (D) and the percentage of on-target and bait-to-bait interactions is shown in (E).

The number of reads which were 'on-target' meaning they contained a baited fragment ranged from 43% to 88% on-target, with an average of 65.6%. Bait-to-bait reads where both fragments were baited represented an average of 12.7% of the reads (Figure 29E).

Invalid di-tags accounted for 9-14% (average of 12.3%) of the total di-tags. The invalid di-tags consisted of self-circularised fragments (1%), the same internal fragments (2%), the same fragment with dangling ends (1%), re-ligated fragments (5%) and wrong sized fragments (5%) (Figure 30).

Figure 30: Average valid and invalid di-tags from HiSeq reads of both captures



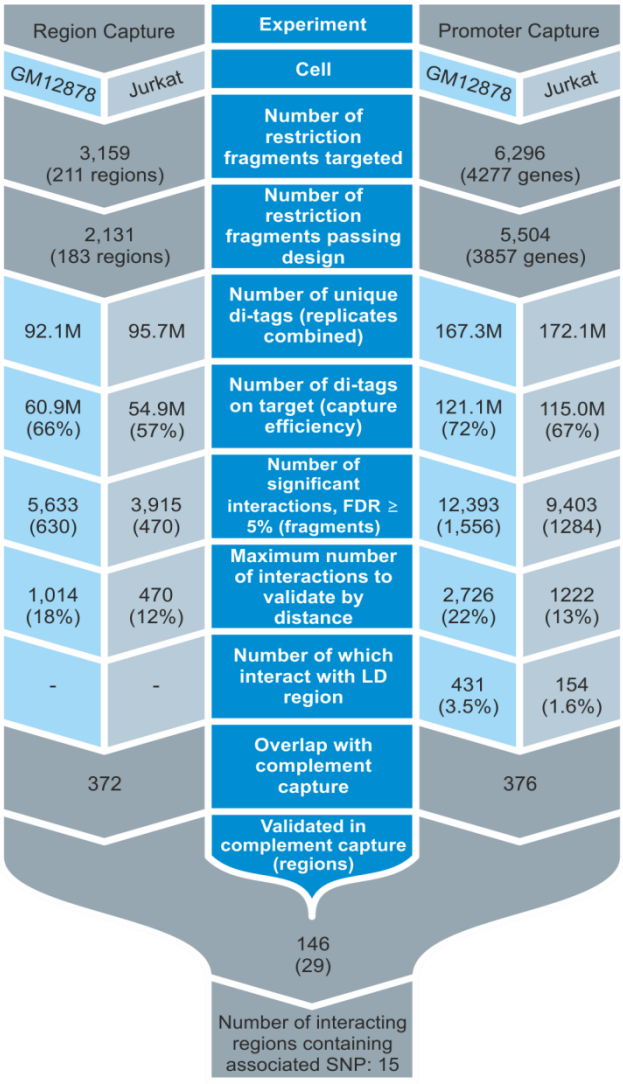
The HiCUP reports generated from the HiSeq sequencing data was used to produce plots of the summary statistics. The percentage of each type of invalid di-tag is shown in the pie-chart. The percentages shown are from an average of all the sequencing runs.

3.4. Significant interactions from Capture Hi-C

Significant interactions were assigned as those interactions seen at a frequency above expected background after biological replicates were pooled, adjusting for distance and mappability, and using an FDR of ≤5% (see section 2.4.4.2), after fitting a negative binomial distribution to the data. 'Confirmed interactions' were those seen in both capture experiments. Other interactions were present in one of the captures and were therefore not validated.

Figure 31 summarises some of the key statistics from the capture experiments. To briefly summarise the CHi-C data, 8594 interactions representing 764 *Hind*III fragments were identified in the Region Capture experiment. Out of 116 restriction fragments, 4.3% contained interactions involving a promoter within 500kb and could be independently validated in the Region Capture experiment. Independent validation implicated 29 disease associated regions, 15 of which contained GWAS SNPs. Approximately 20% of the interactions occurred in both cell lines (Martin *et al.* 2015).

Figure 31: Summary of the CHi-C experiments



Summary statistics from the capture experiments are shown in Table 15, providing information about the average region size, number of interactions per restriction fragment, interaction distances, number of regions/fragments interacting with a promoter, number of SNPs in the interacting regions/fragments, average distance of promoters to region and the number of genes showing interactions.

Table 15: Summary statistics from the capture experiments

		Cell	
Experiment		**GM12878**	**Jurkat**
Region capture	Average Region Size (bp)	64,787	
	Number of *Hin*dIII Fragments	2,131 (183 regions)	
	Average Number of Interactions per Region	48.7	38.2
	Average Number of Interactions per Restriction Fragment	8.9	8.3
	Average Interaction Distance (bp)	1,451,825	1,430,506
	Number of Regions Interacting with a Promoter	37	25
	Number of Fragments Interacting with a Promoter	136	63
	Number of Associated SNPs in the Interacting Regions (r^2≥0.8)	5,329	
	Number of Associated SNPs in the Interacting Fragments (r^2≥0.8)	671	449
Promoter Capture	Average Number of Genes in 1Mb Regions	359	
	Number of *Hin*dIII Fragments	5,504 (3857 genes)	
	Average Distance of Promoters to Region	212,762	
	Number of Genes that show Interactions	1,341	1,136
	Average Interaction Distance (bp)	1,262,173	1,472,446

Filtered data was analysed in-house by several members of the group using published datasets. The RA Immunochip (Eyre *et al.* 2012) and RA trans-ethnic GWAS meta-analysis (Okada *et al.* 2014) datasets were used, along with Immunochip datasets for JIA/PsA. Each disease associated region included in the capture experiments was visualised separately using the WASHU genome browser in order to identify long-range interactions overlaying disease associated SNPs and/or genes.

Due to the scale of the analysis, members of the group were allocated a set of chromosomes to analyse. A table was compiled (analysis of chromosomes 3-6 – see Appendix Table 51), containing information about which dataset was analysed, the index SNP, if any interactions overlayed the index SNP or if any interactions involved disease associated genes. Interactions which were observed in both region capture and promoter capture experiments, in multiple datasets were put

forward as potential interactions to follow-up (Table 16). Disease-associated regions which would be interesting to follow up in future experiments, such as those containing long-range interactions implicating novel candidate genes or interactions involving multiple loci were identified and are summarised in Table 17.

Table 16: Summary of regions containing long-range interactions involving SNPs and/or disease associated genes

Region	Index SNP	Potential interactions for follow-up
1q32	rs17668708	*PTPRC – NEK7*
2p14	rs1858037	*SPRED2*
2q32	rs11889341	*STAT4* (3' to 5' interaction)
3p24.1	rs3806624	*EOMES* interaction with *AZI2*
4p15	rs932036, rs11933540	*RBPJ – STIM2*
4		*ELMO1*
5q11.2	rs7731626	*ANKRD55* with *IL6ST, IL31RA, DDX5*
6q23	rs6920220,rs7752903,rs17264332,rs610604	*IL22RA* to beyond *TNFAIP3*
7		*CDK6*
7p15.2	rs67250450, rs10260837	*HOX – HOTTIP*
10		*GATA3*
10q21	rs12764378, rs71508903	*ARID5B - RTKN2* and *ARID5B-ARID5B* 5'-3'
10p15.1	rs706778, rs10795781, rs947474	*IL2RA*
11p12	rs331463	*TRAF6*
12		*CD6 – CD5, CCDC86*
13		*CUL5 – RDX*
14	rs1950807, rs12434551, rs3825568	*FOXO1 – COG6*
15		*RAD51B*
16p13.13	rs4780471, rs12928822	*CLEC16A/DEXI*
17		*PRKCH – HIF1A*
20		*SPATA2* (PSA)
21		*AIRE*
21	rs9979383	*RUNX1*
22		*IL2RB*

Table 17: Summary of the regions chosen for follow-up studies

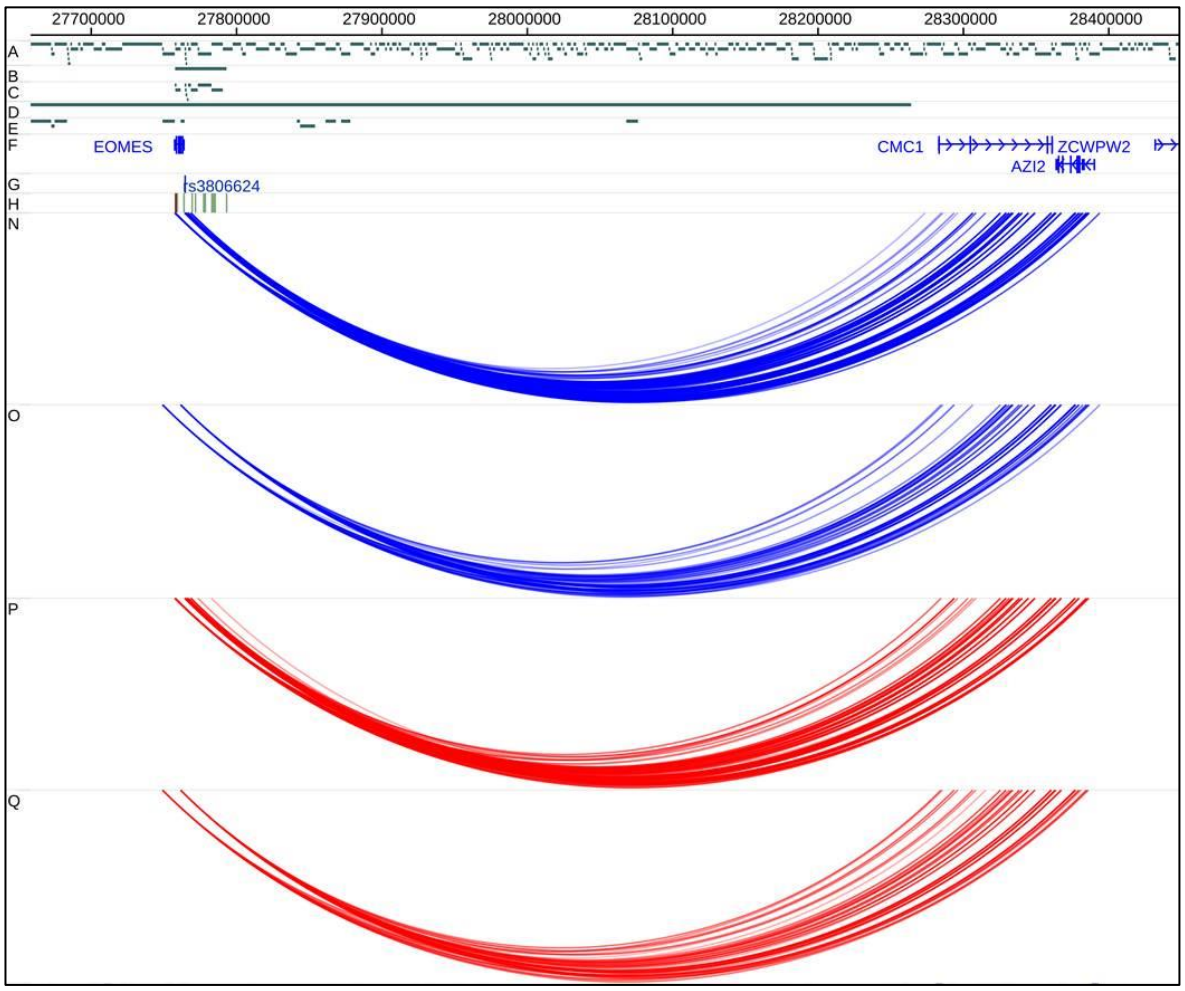
Region	Associated SNPs	Interesting interaction	Implicated disease
1q32	rs17668708, rs2477077, rs2014863	*PTPRC – DENND1B*	**PsA and RA**
2q32	rs11889341	*STAT4*	**RA/JIA**
3p24	rs3806624	*EOMES-AZI2*	**RA**
4p15	rs932036, rs11933540	*RBPJ* – lncRNA	**RA, T1D**
5q11	rs6859219	*ANKRD55 – IL6ST, IL31RA*	**RA**
6q23	**rs6920220,rs7752903,rs17264332,rs610604**	***IL22RA* to beyond *TNFAIP3***	**RA/PsA**
7p15.2	rs67250450, rs10260837	*HOX – HOTTIP*	**RA**
10q21	rs12764378, rs71508903	*ARID5B - RTKN2*	**RA**
10p15.1	rs706778, rs10795781, rs947474	*IL2RA-PRKCQ*	
11p12	rs331463	*TRAF6*	**RA**
Chr13	rs7993214	*COG6-FOXO1*	**RA/JIA**
Chr14	rs1950807, rs12434551, rs3825568	*RAD51B – ZFP36L1*	**RA**
16p13.13	rs4780471, rs12928822	*CLEC16A/DEXI*	**RA, PsA and T1D**
21q22.12	rs9979383	*RUNX1*	**RA, JIA**

Some of the significant findings from the analysis are summarised below and are described in further detail in the manuscript included in Appendix 2 (Martin *et al.* 2015). An additional WASHU plot showing *IL2RA* interactions is shown in Appendix 1 Figure 67.

a) Interactions implicating novel candidate genes

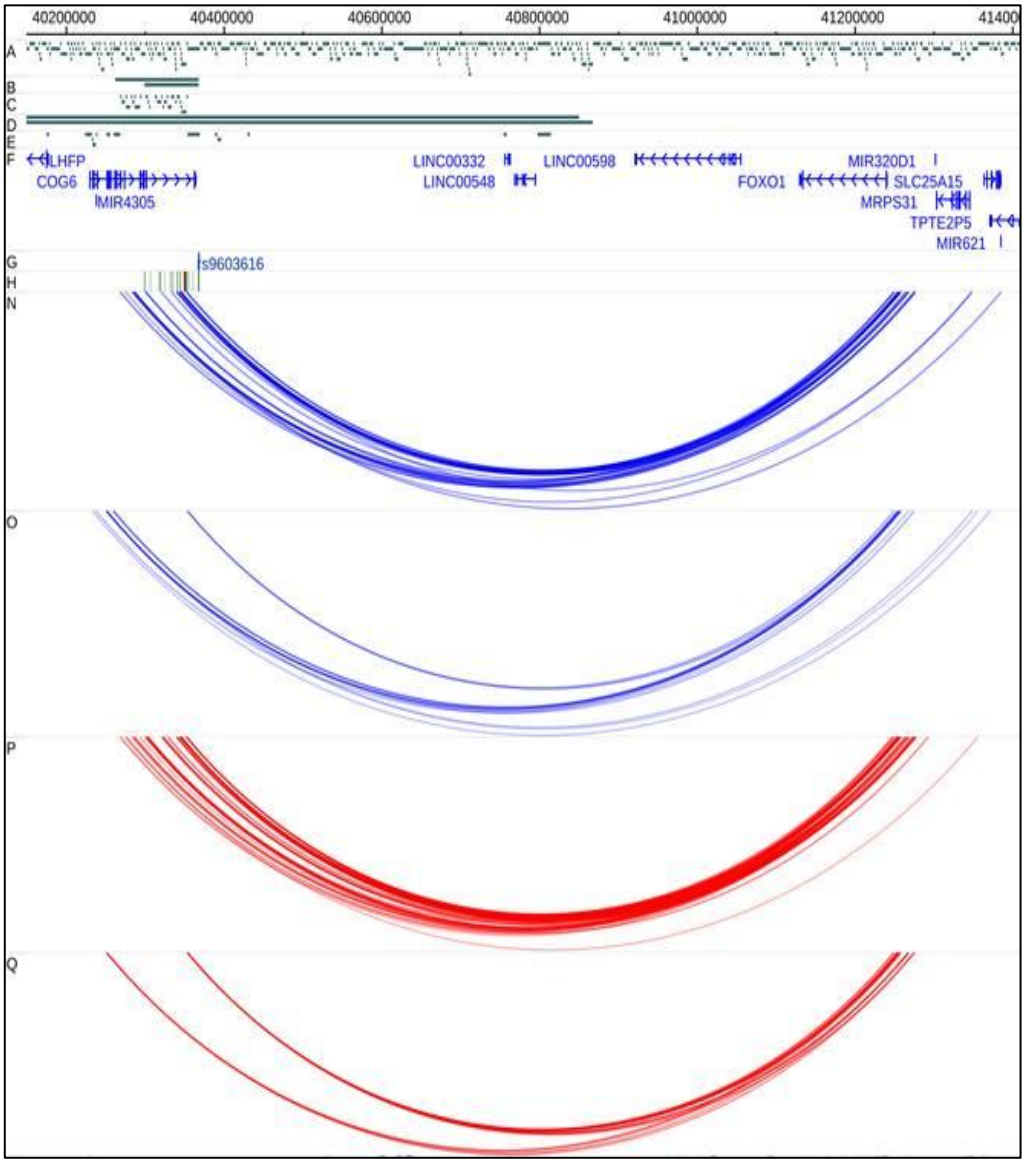
Disease-associated SNPs were found to often interact not with the nearest gene but with promoters some distance away. Both GM12878 and Jurkat cell lines showed that SNPs situated proximal to the *EOMES* gene (involved in differentiation of effector CD8+ T-cells) in the 3p24 region interacted with the promoter of *AZI2I* (involved in NF-κB activation) situated ~640kb away (Figure 32). Also, variants associated with RA and juvenile idiopathic arthritis (JIA) in the 3' intronic region of *COG6* (encoding a component of Golgi apparatus) interacted with the *FOXO1* promoter, over 1Mb away, in both cell types (Figure 33).

Figure 32: Long-range interaction between *EOMES* and *AZI2I*



Genomic co-ordinates are shown along the top of each panel and tracks are labelled (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) Index SNPs identified for RA (H) Density plots showing 1000 Genomes SNPs in LD (r^2>0.8) with the index SNPs for RA Immunochip study; (N–Q) Significant Interactions identified in the region and promoter capture experiments in GM12878 (N,O) and Jurkat (P,Q) cells.

Figure 33: Long-range interaction between *COG6* and *FOXO1*

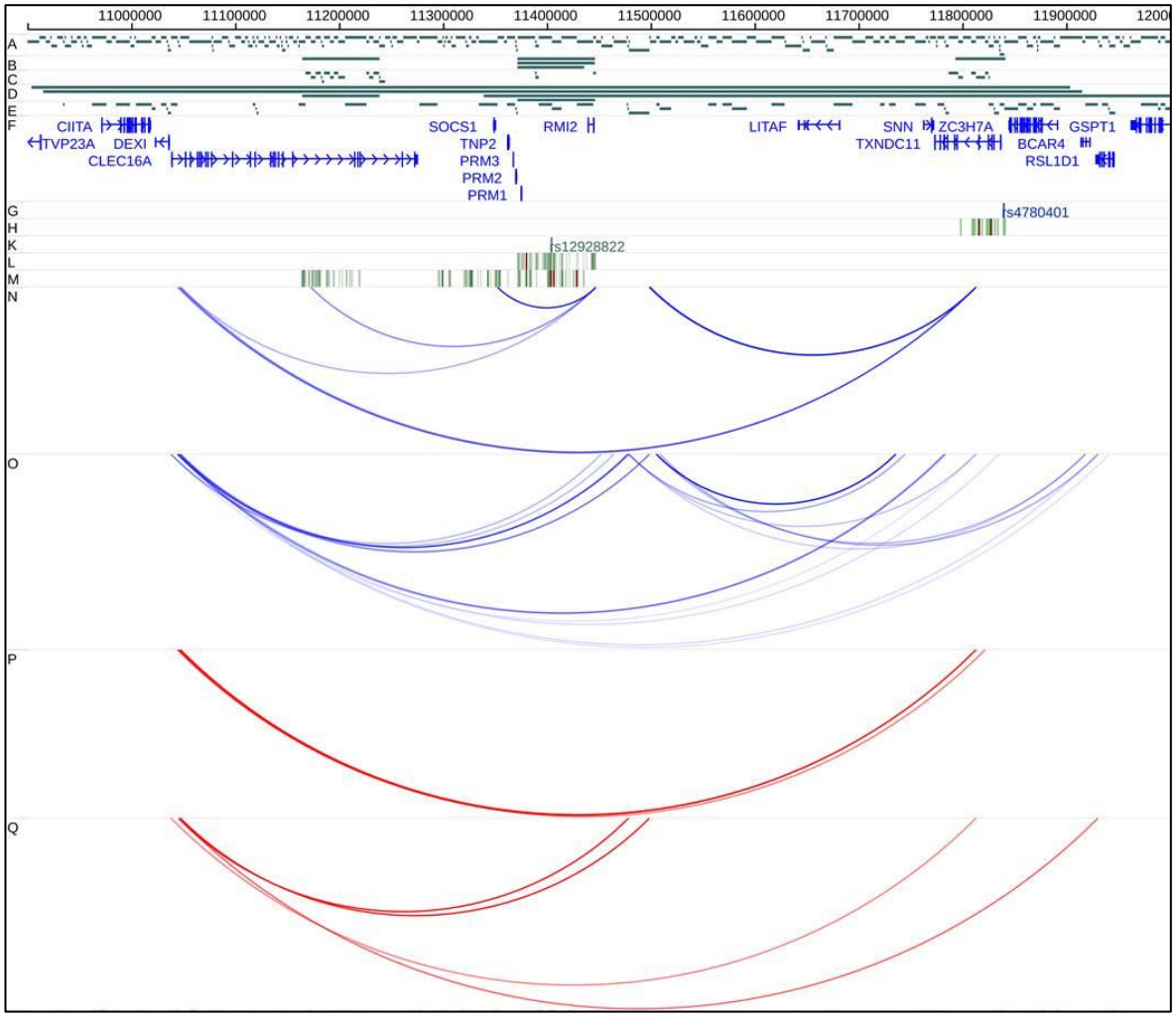


Genomic co-ordinates are shown along the top of each panel and tracks are labelled (A) HindIII restriction fragments; (B–E) Regions targeted and restriction fragments included in the region (B,C) and promoter (D,E) capture experiments; (F) RefSeq Genes from the UCSC Genome Browser; (G) Index SNPs identified for RA (H) Density plots showing 1000 Genomes SNPs in LD (r^2>0.8) with the index SNPs for RA Immunochip study; (N–Q) Significant Interactions identified in the region and promoter capture experiments in GM12878 (N,O) and Jurkat (P,Q) cells.

b) Interactions involving multiple genetic risk loci have common interaction targets

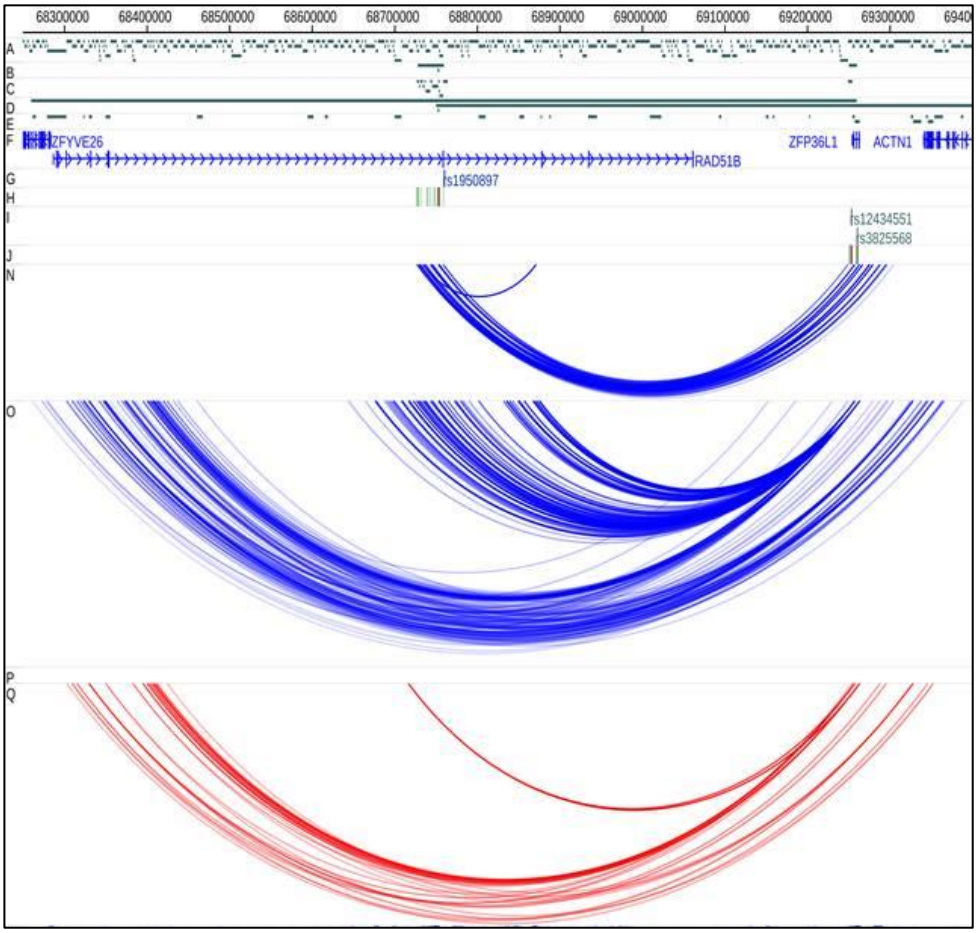
Genetic regions containing susceptibility loci for different autoimmune diseases, mapping some distance apart, were found to interact with a common target. In both GM12878 and Jurkat cell lines 16p13 SNPs associated independently with RA, psoriatic arthritis (PsA) and T1D interacted with the *DEXI* promoter and in addition, RA and JIA SNP regions interacted with each other in GM12878 cells (Figure 34). RA associated variants located within a strong enhancer of *RAD51B* interacted with the promoter of *ZFP36L1*, a zinc finger transcription factor involved in the transition of B-cells to plasma cells, which also contains SNPs associated with JIA (Figure 35). Variants associated with PsA, within *DENND1B* were shown to interact with *PTPRC,* a region independently associated with RA (Figure 36).

Figure 34: RA, PsA and T1D SNPs all interact with the *DEXI* promoter



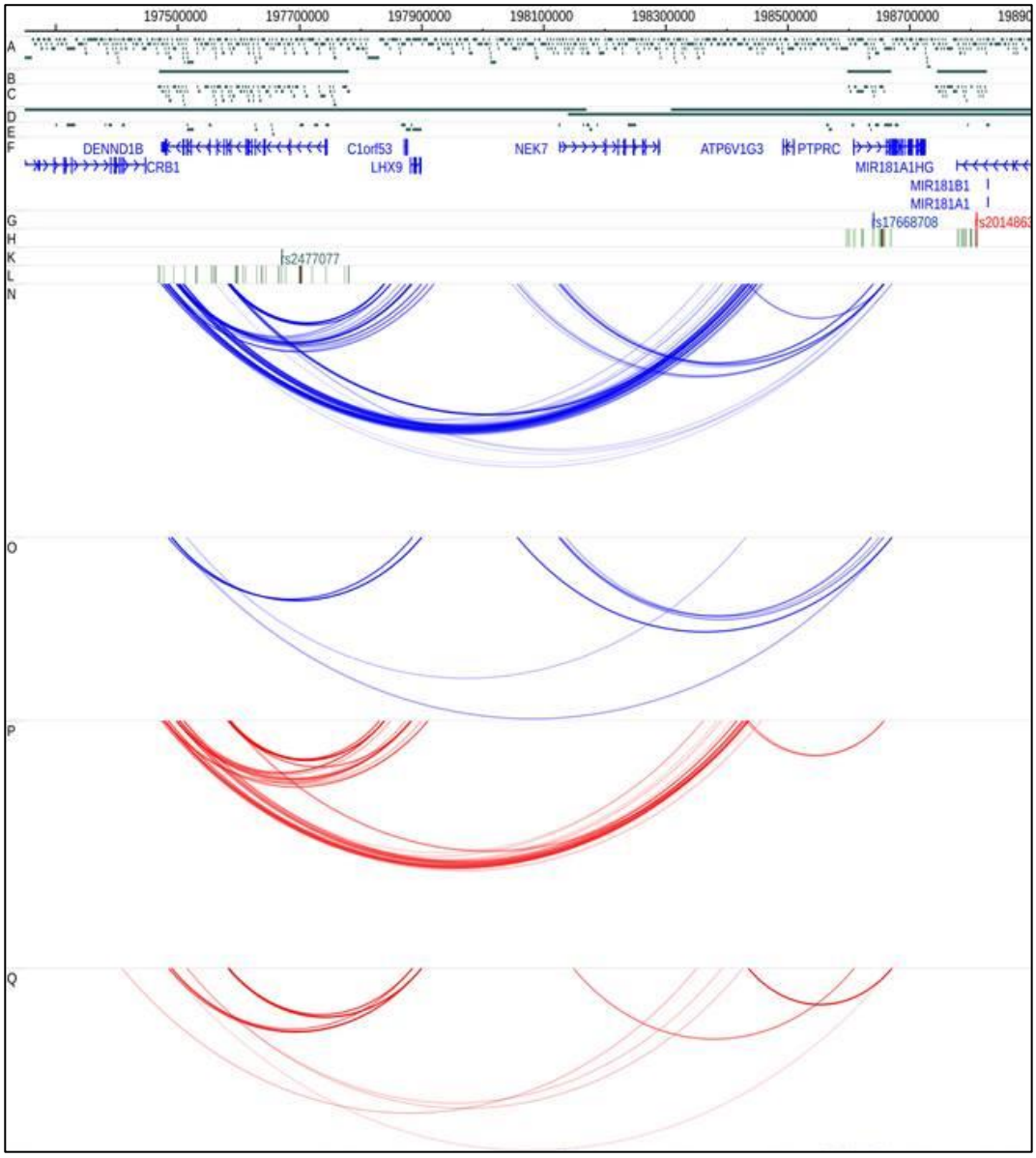
Genomic co-ordinates are shown along the top of each panel. (A) HindIII restriction fragments; (B–E) Regions targeted and restriction fragments included in the region (B,C) and promoter (D,E) capture experiments; (F) RefSeq Genes from the UCSC Genome Browser; (G) Index SNPs identified for RA, (K) PsA. Density plots showing 1000 Genomes SNPs in LD (r^2>0.8) with the index SNPs (green–red) for RA (H), 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.

Figure 35: RA and JIA SNPs implicate both *ZFP36L1* and *RAD51B*



Genomic co-ordinates are shown along the top. (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; (G) Index SNPs identified for RA and (I) JIA; Density plots showing 1000 Genomes SNPs in LD (r^2>0.8) with the index SNPs for RA (H) and JIA (J); (N–Q) Significant Interactions identified in the region and promoter capture experiments in GM12878 (N,O) and Jurkat (P,Q) cells.

Figure 36: PsA variants within *DENND1B* interact with *PTPRC*, independently associated with RA

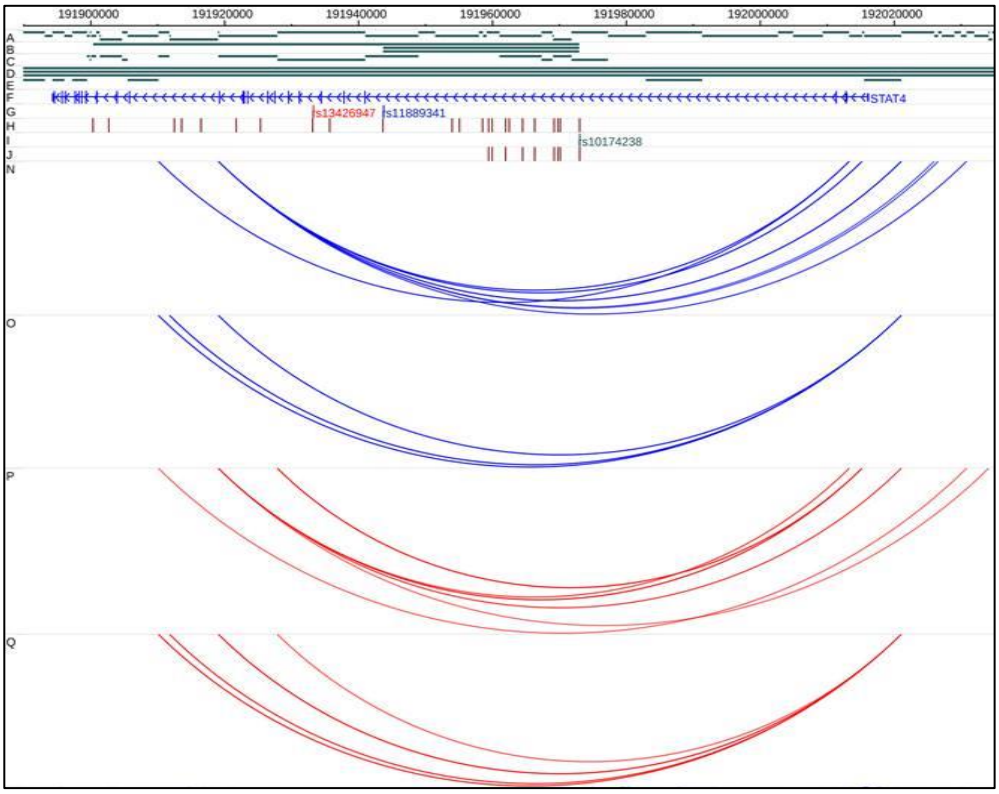


Genomic co-ordinates are shown along the top. (A) HindIII restriction fragments; (B–E) Regions targeted and restriction fragments included in the region (B,C) and promoter (D,E) capture experiments; (F) RefSeq Genes from the UCSC Genome Browser; (G) Index SNPs identified for RA and (K) PsA.; Density plots showing 1000 Genomes SNPs in LD (r^2>0.8) with the index SNPs for RA (H) and PsA (L); (N–Q) Significant Interactions identified in the region and promoter capture experiments in GM12878 (N,O) and Jurkat (P,Q) cells.

c) Interactions with loci previously implicated in disease

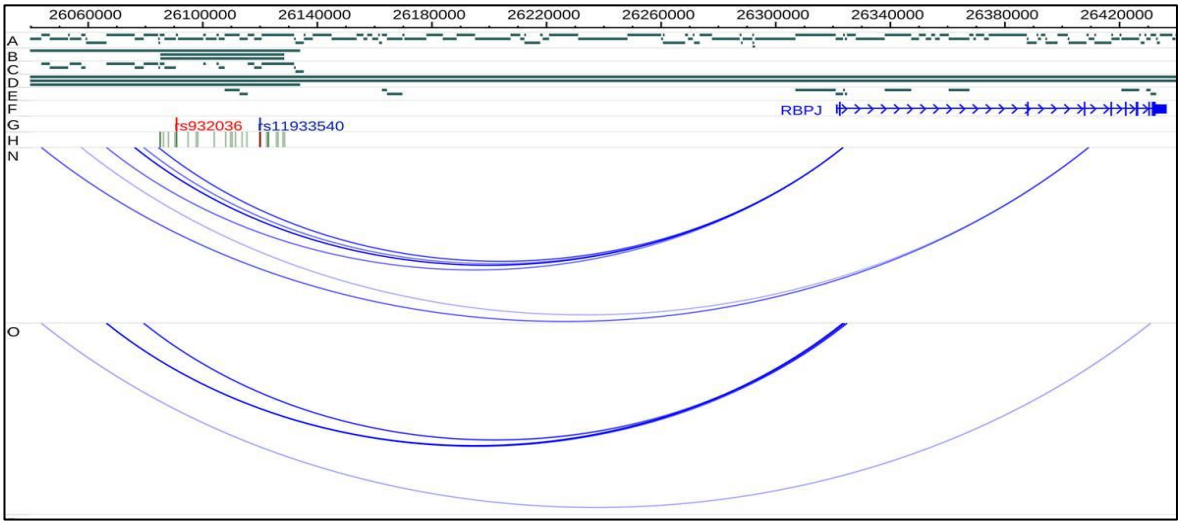
Loci which had previously shown evidence as being associated with RA in GWAS and fine-mapping studies showed a number of interesting interactions. *STAT4* intronic SNPs associated with RA and JIA were found to interact with the *STAT4* promoter (Figure 37). Associated SNPs within a lncRNA interacted with the *RBPJ* gene promoter in GM12878 cells (Figure 38). SNPs located within an intron of *ARID5B,* interacted with the promoter of *ARID5B* and also displayed a long range interaction with *RTKN2* (Figure 39).

Figure 37: Intronic SNPs within *STAT4* interact with the *STAT4* promoter



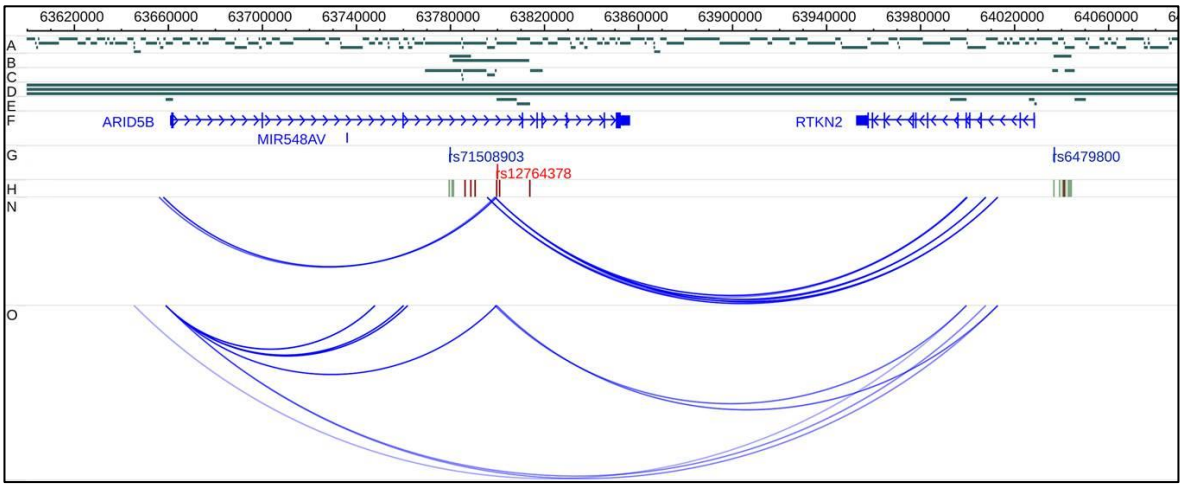
Genomic co-ordinates are shown along the top. (A) HindIII restriction fragments; (B–E) Regions targeted and restriction fragments included in the region (B,C) and promoter (D,E) capture experiments; (F) RefSeq Genes from the UCSC Genome Browser; (G) Index SNPs identified for RA and (I) JIA; Density plots showing 1000 Genomes SNPs in LD (r^2>0.8) with the index SNPs for RA (H) and JIA (J); (N–Q) Significant Interactions identified in the region and promoter capture experiments in GM12878 (N,O) and Jurkat (P,Q) cells

Figure 38: Associated SNPs within a lncRNA interact with the *RBPJ* promoter



Genomic co-ordinates are shown along the top. (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; (G) Index SNPs identified for RA; Density plot showing 1000 Genomes SNPs in LD (r^2>0.8) with the index SNPs for RA (H); (N–Q) Significant Interactions identified in the region and promoter capture experiments in GM12878 (N,O) cells.

Figure 39: SNPs located within an intron of *ARID5B*, interacted with the *ARID5B* promoter and also with *RTKN2*



Genomic co-ordinates are shown along the top. (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; (G) Index SNPs identified for RA; Density plot showing 1000 Genomes SNPs in LD (r^2>0.8) with the index SNPs for RA (H); (N–Q) Significant Interactions identified in the region and promoter capture experiments in GM12878 (N,O) cells.

3.5. The 6q23 locus and a new candidate causal gene?

The region that was chosen for further investigation was the 6q23 locus which is an important locus in autoimmunity and has been implicated in multiple diseases by GWAS, where independent variants have been found to be associated with different autoimmune diseases (Figure 40). The region capture experiment, targeted the LD blocks (r^2>0.8) containing SNPs associated with autoimmune disease - rs6920220 (RA, T1D, JIA), rs7752903 (RA) and rs610604 (psoriasis and PsA). The promoter capture experiment targeted all known gene promoters overlapping the region 500kb up and downstream of the lead disease associated SNPs.

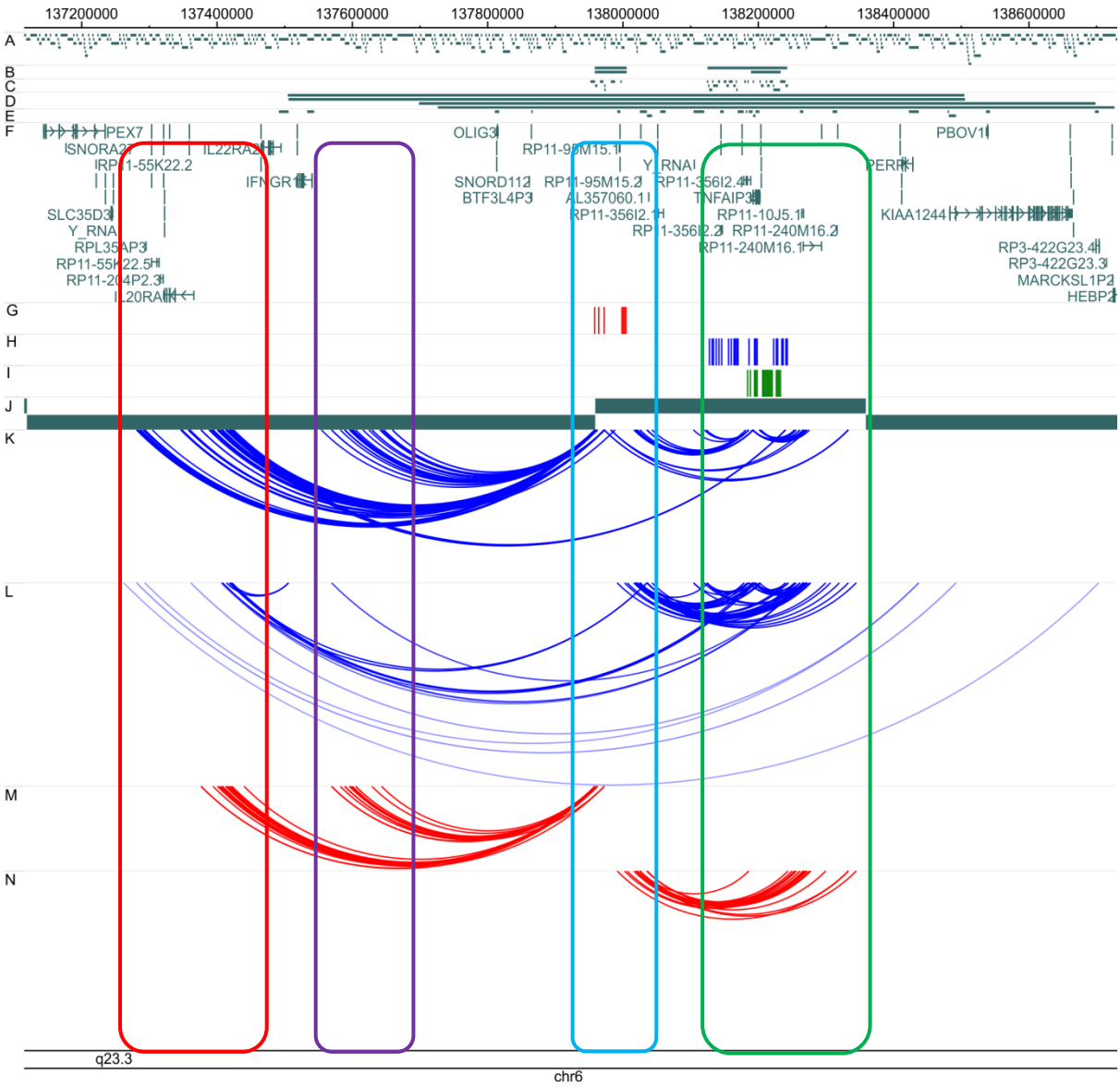
The LD block containing rs6920220 spans 47.3kb (chr6:137959235-138006504) and contains seven restriction fragments (highlighted with a blue block). Five out of the seven restriction fragments were involved in a complex pattern of statistically significant long-range interactions. These interactions involved genes (*IL20RA* and *IFNGR1*) (highlighted with a red block and purple block respectively), and lncRNAs downstream of the *TNFAIP3* gene (RP11-10J5.1 and RP11-240M16.1) (highlighted with a green block).

The region capture experiment targeting the LD block containing rs7752903 (RA) and rs610604 (PsA) spanned the *TNFAIP3* gene and the regions upstream and downstream. Interactions were identified with a region proximal to the rs6920220 LD block, encompassing the lncRNAs RP11-95M15.2 (a *PTPN11* pseudogene) and RP11-356I2.1, and the miRNA AL357060.1 and also an upstream region containing non-coding RNAs (Y_RNA and RP11-356I2.2). Interactions were also detected which involved the *TNFAIP3* gene and the downstream lncRNAs RP11-10J5.1 and RP11-240M16.1. These same lncRNAs interacted with the rs6920220 LD block, and with the *IL20RA* gene.

The promoter capture experiment independently validated the interactions identified in the region capture and also identified an interaction between the promoters of *TNFAIP3* and *IL20RA* that could not be detected in the region capture due to the exclusion of promoters in that capture design.

The co-ordinates of the interacting fragments were obtained from the WASHU genome browser (Appendix Table 52) and these fragments formed the basis for an in-depth 3C-qPCR analysis of the interacting regions.

Figure 40: Capture Hi-C identifies long range interactions in the 6q23 locus



Long-range interactions were visualised using the WASHU genome browser. Genomic co-ordinates are shown along the top of the panel and tracks are labelled A-N: A – *Hin*dIII 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, H, I, –1000 Genomes SNPs in LD (r^2≥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); K-N – Significant Interactions identified in the Region and Promoter capture experiments in GM12878 (K, L) and Jurkat (M, N) cells.

3.6. Bioinformatic analysis of the 6q23 locus

RA associated SNPs in the 6q23 intergenic region between the *TNFAIP3* and *OLIG3* genes were shown in the Capture Hi-C experiments to be involved in long-range interactions with the *IL20RA* and *IFNGR1* genes and also with lncRNAs which can be involved in gene regulation. Bioinformatic analysis of the lead RA associated SNP rs6920220 was carried out to pinpoint SNPs with regulatory potential in order to prioritise the most plausible causal SNP.

RegulomeDB was also used to investigate the functional potential of the RA SNPs and showed that, of the SNPs in LD with rs6920220, rs6927172 scored 2b (likely to affect binding) and rs35926684 scored 3a (less likely to affect binding) (Table 18). The other SNPs in LD scored 6 (minimal binding evidence). Two SNPs in LD with rs6920220 (rs6933404 and rs11757201) did not have any data available in RegulomeDB so it is unclear as to their potential functional effect. The UCSC track from RegulomeDB for rs6927172 (Figure 41A) shows that the SNP is in a DNase cluster, has transcription factor binding and is in a conserved region, which suggests that the SNP is likely to have a functional effect. Additional UCSC tracks from two assemblies, 2006 (NCBI36/hg18) and 2009 (GRCh37/hg19), are also shown (Figure 41B and C), providing further evidence that rs6927172 lies in a potential regulatory region (transcription factor binding, DNase1 hypersensitivity, H3K4me1 enrichment).

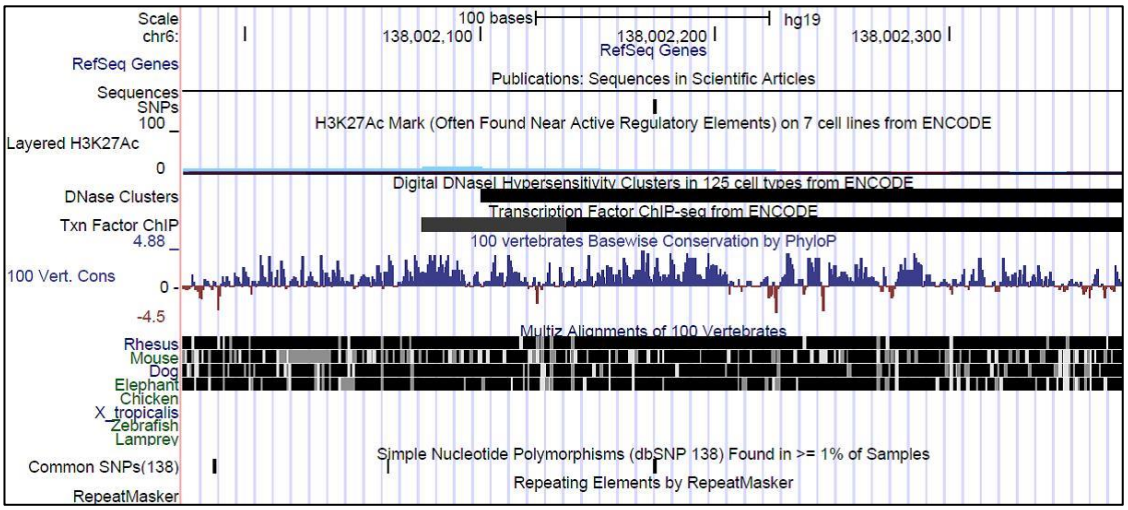
Haploreg v4.1 was used to identify SNPs in LD with rs6920220 (Table 19). Eight SNPs were in strong LD with the associated SNP. Three SNPs had an r^2=1 (rs6927172, rs11757201 and rs17264332), and five SNPs had an r^2>0.8 (rs6933404, rs62432712, rs2327832, rs928722, and rs35926684). Haploreg v4.1 also showed that the rs6927172 SNP had the most evidence of potential regulatory activity. Analysis of chromatin state (ChromHMM and DNase hypersensitivity) showed that rs6927172 mapped to an enhancer region in B-lymphoblasts, T_H17 T-cells and T_{REG} T-cells and mapped to a region of open chromatin. Transcription factor binding sites were present, including NF-κB and BCL3.

Table 18: Results from Regulome DB

Co-ordinates	dbSNP ID	RegulomeDB Score	Functional effect
chr6:138002174	**rs6927172**	**2b**	**Likely to affect binding**
chr6:137999562	rs35926684	3a	Less likely to affect binding
chr6:137964696	rs62462712	6	Minimal binding evidence
chr6:137973067	rs2327832	6	Minimal binding evidence
chr6:137973831	rs928722	6	Minimal binding evidence
chr6:138005514	rs17264332	6	Minimal binding evidence
chr6:138006503	rs6920220	6	Minimal binding evidence
chr6:137959234	rs6933404	No data	
chr6:138003821	rs11757201	No data	

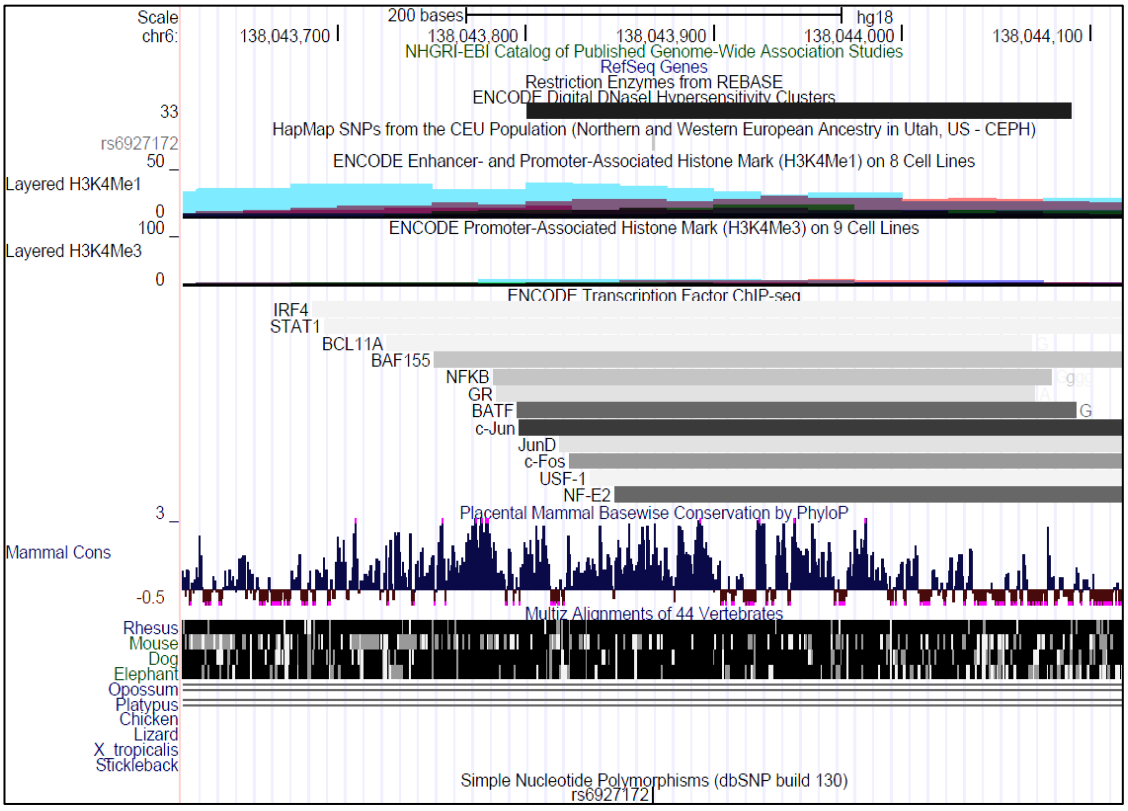
Figure 41: UCSC tracks for the rs6927172 SNP

(A) RegulomeDB Track



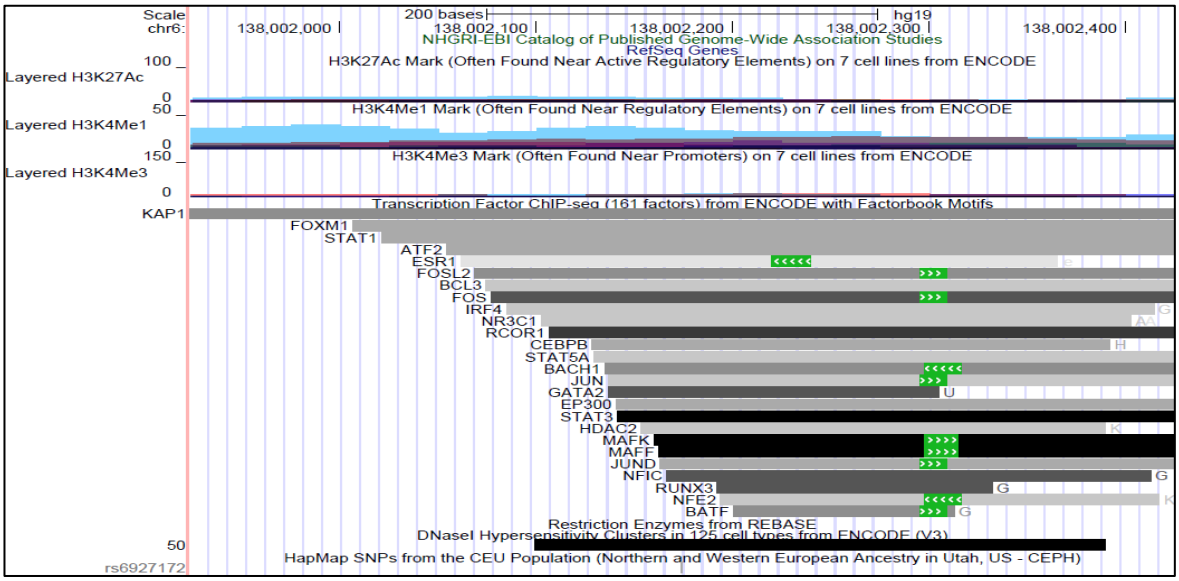
UCSC track generated by Regulome DB showing tracks relating to histone marks, DNase clusters, transcription factor binding and conservation between species.

(B) UCSC Track (2006)



UCSC track from the 2006 (NCBI36/hg18) assembly showing tracks relating to histone marks, DNase clusters, transcription factor binding, conservation between species and location of rs6927172.

(C) UCSC Track (2009)



UCSC track from the 2009 (GRCh37/hg19) assembly showing tracks relating to histone marks, DNase clusters, transcription factor binding and location of rs6927172.

3.6.1. eQTL analysis of proxy SNPs

The role of the proxy SNPs in regulating gene expression was investigated using eQTL databases which search publicly available eQTL data. Genevar eQTL analysis of the lead SNP, rs6920220, using simple linear regression in CEU and all HapMap3 populations showed no significant eQTLs with *TNFAIP3* or any other potential candidate gene (*IL20RA, IFNGR1*) in the 6q23 region covered by the region capture experiment. Gene-centric analysis of genes within the 6q23 region in Genevar using simple linear regression in CEU and all HapMap3 populations also showed no significant eQTLs. Analysis of whole blood using data from the GTEx project also detected no significant association between the rs6920220 SNP and *TNFAIP3*, or any other gene's expression (Figure 42).

Figure 42: Whole blood eQTL analysis plot from GTEx

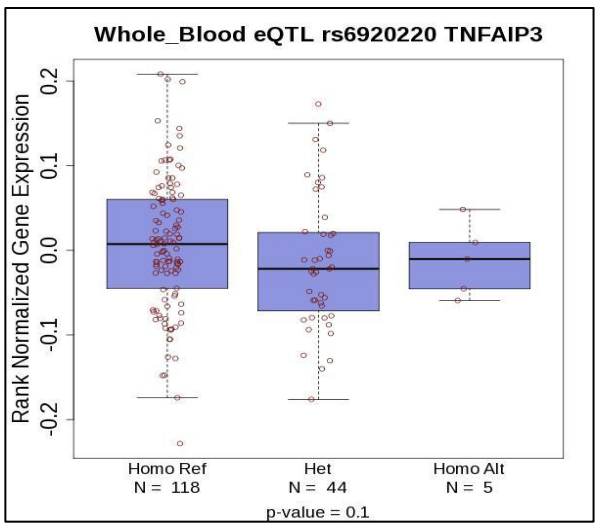


Table 19: Functional annotation of SNPs in the 6q23 intergenic LD block tagged by rs6920220 using Haploreg v4.1

Query SNP: rs6920220 and variants with r^2 >= 0.8

chr	pos (hg38)	LD (r^2)	LD (D')	variant	Ref	Alt	AFR freq	AMR freq	ASN freq	EUR freq	SiPhy cons	Promoter histone marks	Enhancer histone marks	DNAse	Proteins bound	Motifs changed
6	137638098	0.89	0.95	rs6933404	T	C	0.11	0.11	0.00	0.17			BLD, BRN	GI		STAT
6	137643560	0.88	0.94	rs62432712	A	G	0.08	0.10	0.00	0.17						Pax7,RORalpha1,Vax2
6	137651931	0.93	0.97	rs2327832	A	G	0.11	0.11	0.00	0.17			5 tissues	GI,GI,PLCNT		10 altered motifs
6	137652695	0.92	0.96	rs928722	C	T	0.11	0.11	0.00	0.17			5 tissues			4 altered motifs
6	137678425	0.84	0.95	rs35926684	GA	G	0.14	0.12	0.00	0.18			BLD			4 altered motifs
6	137681038	1	1	rs6927172	C	G	0.12	0.10	0.00	0.17		BLD, LNG	7 tissues	14 tissues	13 bound proteins	8 altered motifs
6	137682685	1	1	rs11757201	G	C	0.08	0.10	0.00	0.17						Mrg,Sp4
6	137684378	1	1	rs17264332	A	G	0.12	0.10	0.00	0.17			LNG, BLD			5 altered motifs
6	137685367	1	1	rs6920220	G	A	0.12	0.10	0.00	0.17			BLD			Hltf

Summary of Results Section 1

- Quality control of Capture Hi-C libraries, both pre-capture and post-capture, showed that the samples were of a consistently high standard.

- Analysis of Capture Hi-C data revealed numerous interesting long-range interactions that implicated novel candidate genes and showed that interactions involving multiple genetic loci could have common interaction targets.

- Analysis of the 6q23 region revealed that SNPs associated with RA interacted with novel candidate genes, *IL20RA, IFNGR1* and also with regulatory lncRNAs downstream of *TNFAIP3.*

- Bioinformatic analysis provided evidence that the rs6927172 SNP was the most likely regulatory SNP in the 6q23 intergenic region.

4. Results Section 2

Validation of long-range interactions in the 6q23 locus by 3C-qPCR

4.1. Validation of long-range interactions in the 6q23 locus by 3C-qPCR

3C libraries were prepared from HapMap B-cell lines specific for the appropriate genotype. Libraries were also prepared from Jurkat T-cells and from primary human synovial fibroblasts (provided by Dr. Caroline Ospelt, University Hospital Zurich, Switzerland). All 3C libraries were prepared using the crosslinking, digestion, and ligation steps of the Hi-C protocol excluding biotin dATP fill-in (Section 2.3.2).

A control 3C template was generated using minimally overlapping BAC clones (Children's Hospital Oakland Research Institute; Life Technologies) spanning the region encompassing *IL20RA* and lncRNAs downstream of *TNFAIP3* (Miele et al. 2006) (Chr6:137286536-138591433) (Appendix Figure 65).

All qPCR was performed in triplicate using Power SYBR green (Life Technologies) on a QuantStudio 12K Flex instrument (Life Tech). For each set of interactions a standard curve was generated using the BAC control libraries, 50ng of 3C library was used per reaction and a no-template control was included. Relative interaction frequency was calculated as described in section 2.6.4. (Hagege *et al.* 2007).

4.1.1. RA 6q23 SNP interactions

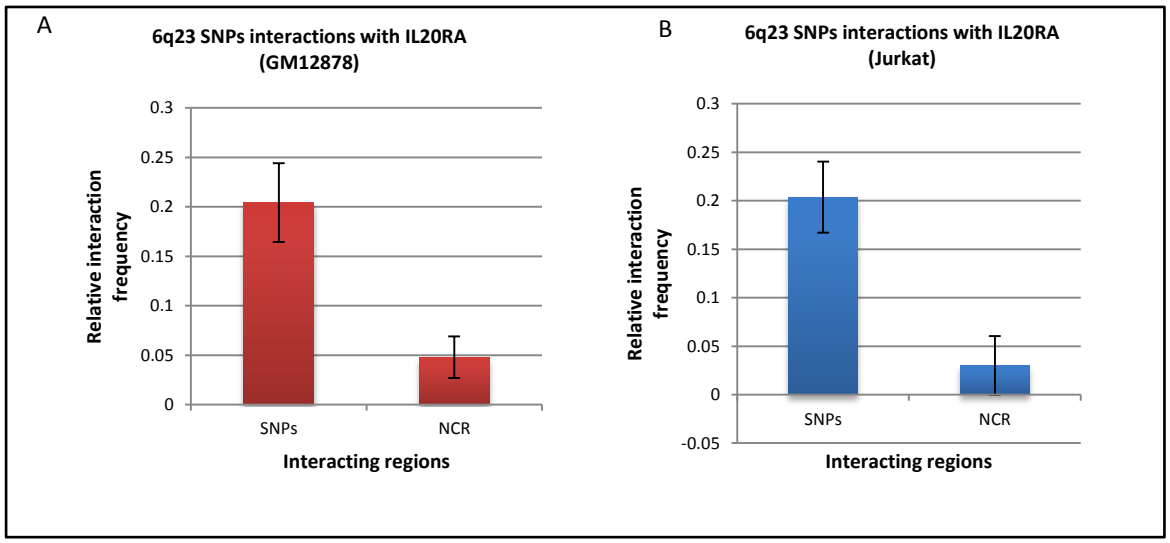
1. *IL20RA* promoter region interacts with 6q23 RA SNPs in T-cells, primary human synovial fibroblasts and in a genotype-specific manner in B-cells

An interaction was observed between a *Hin*dIII fragment located in the *IL20RA* promoter region (IL20RA_2 chr6:137403878-137407040) and a *Hin*dIII fragment located in an LD block containing RA associated SNPs (SNPs_1 chr6:137952897-137959707). This interaction was present in both B-cell and T-cell lines, with a 1% FDR and only in the region capture experiment since *IL20RA* maps around 680kb from rs6920220, outside the boundaries of the region targeted by the promoter capture.

The SNPs_1 fragment is located at the 5' region of the RA LD block and showed a low level of interaction in initial 3C-qPCR assays but was significantly increased over the negative control region in both GM12878 B-LCLs (p=0.023) and Jurkat T-cells (p=0.039) (Figure 43A and B). However, 3C-qPCR using primers spanning the SNPs LD block showed that the interaction peak localised to a *Hin*dIII fragment downstream of the SNPs_1 fragment containing RA associated SNPs rs10499194 and rs6927172 which is in perfect LD with the lead RA SNP rs6920220 and has been predicted through bioinformatics to be most likely to have a functional effect (Figure 44). Assays were performed using 3C libraries generated from HapMap B-LCLs specific for the appropriate SNP genotype - GM12878, GM12145 (CG); GM11993, GM10838 (CC); GM07037, GM10850, GM10858 (GG) (Figure 44A). The interaction with the rs6927172 SNP fragment occurred more frequently in LCLs containing the risk allele (G) compared to the LCLs that were

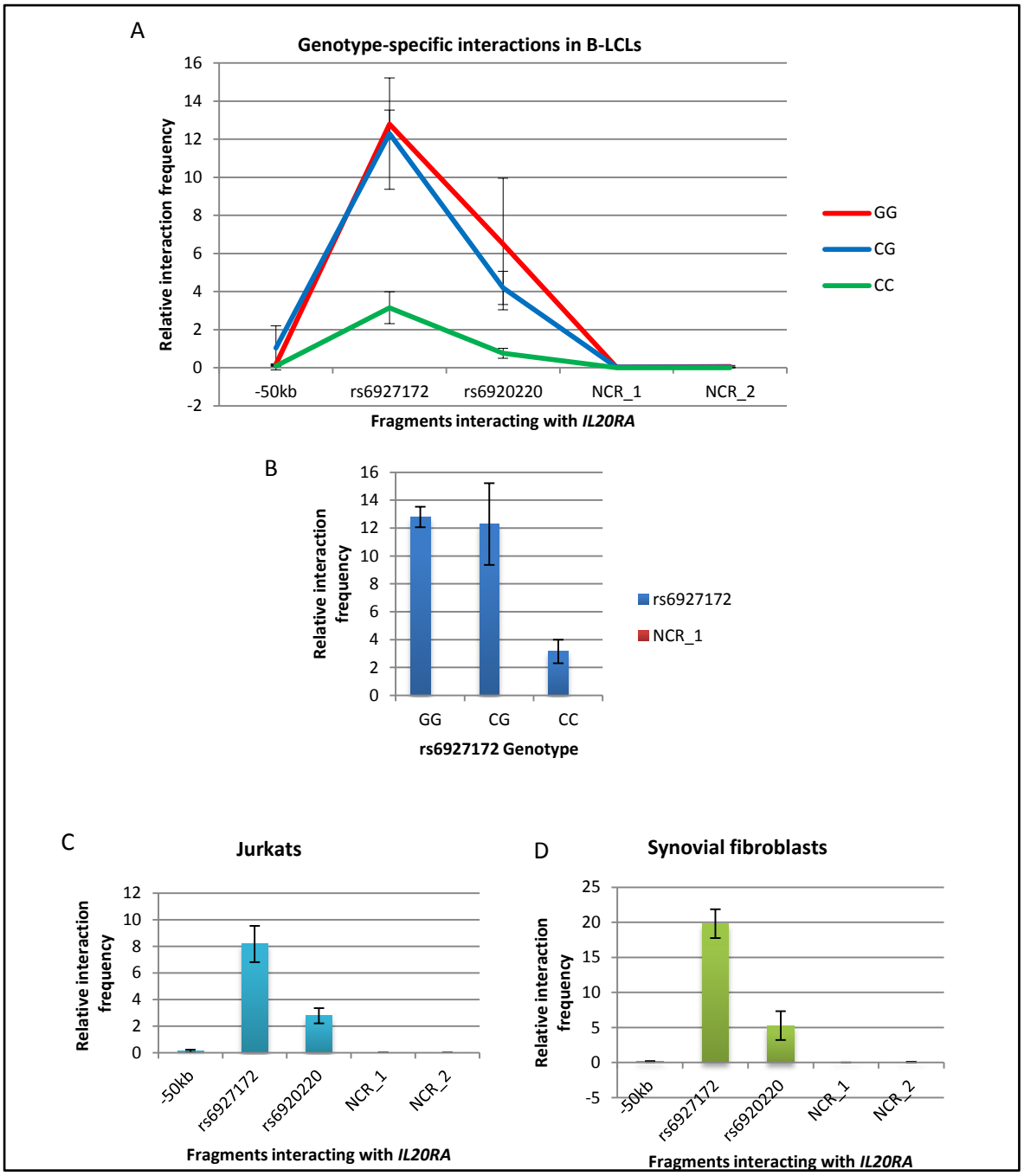
homozygous for the non-risk (C) allele (p=0.034) (Figure 44B). Jurkat T-cells (Figure 44C) and primary human synovial fibroblasts (Figure 44D) were also analysed, with the interaction peak localising to the same fragment.

Figure 43: Initial 3C-qPCR analysis of *IL20RA* interactions with the 6q23 SNP region



3C-qPCR was carried out using the anchor primer IL20RA_2B in combination with a test primer (SNPs_1) designed for the interacting *Hin*dIII fragment or a non-interacting region (NCR). For each set of primers a standard curve was generated using BAC control libraries spanning the region of interest. SYBR green qPCR was carried out in triplicate using 50ng of 3C library per reaction or a no-template (water) control. Triplicate 3C libraries were analysed (BR = biological replicate): A = GM12878, B = Jurkat.

Figure 44: Interactions between *IL20RA* and the 6q23 SNP region



3C-qPCR was carried out using the anchor primer IL20RA_2B in combination with primers designed in multiple *Hin*dIII fragments within the RA SNPs LD block - SNPs_1 (-50kb), a *Hin*dIII fragment containing the rs6920220 RA associated SNP, a *Hin*dIII fragment containing a putative functional RA SNP rs6927172, along with two NCRs. For each set of primers (anchor fragment primer + test region or NCR) a standard curve was generated using BAC control libraries spanning the region of interest. SYBR green qPCR was carried out in triplicate using 50ng of 3C library per reaction or a no-template control. Assays were performed using 3C libraries generated from HapMap B-LCLs specific for the appropriate rs6927172 SNP genotype - GM12878, GM12145 (CG); GM11993, GM10838 (CC); GM07037, GM10850, GM10858 (GG).

The data represents the average interaction frequencies of the samples tested; error bars are +/- Standard Deviation (St.Dev). T-tests were used to compare samples homozygous for the risk allele (CC) and samples homozygous for the non-risk (GG) allele (p=0.034). (A) Genotype-specific interactions in B-LCLs at the different *Hin*dIII fragments, (B) Genotype-specific interactions at the *Hin*dIII fragment containing the rs6927172 SNP, (C) Interactions in Jurkats, (D) Interactions in synovial fibroblasts (only two libraries available).

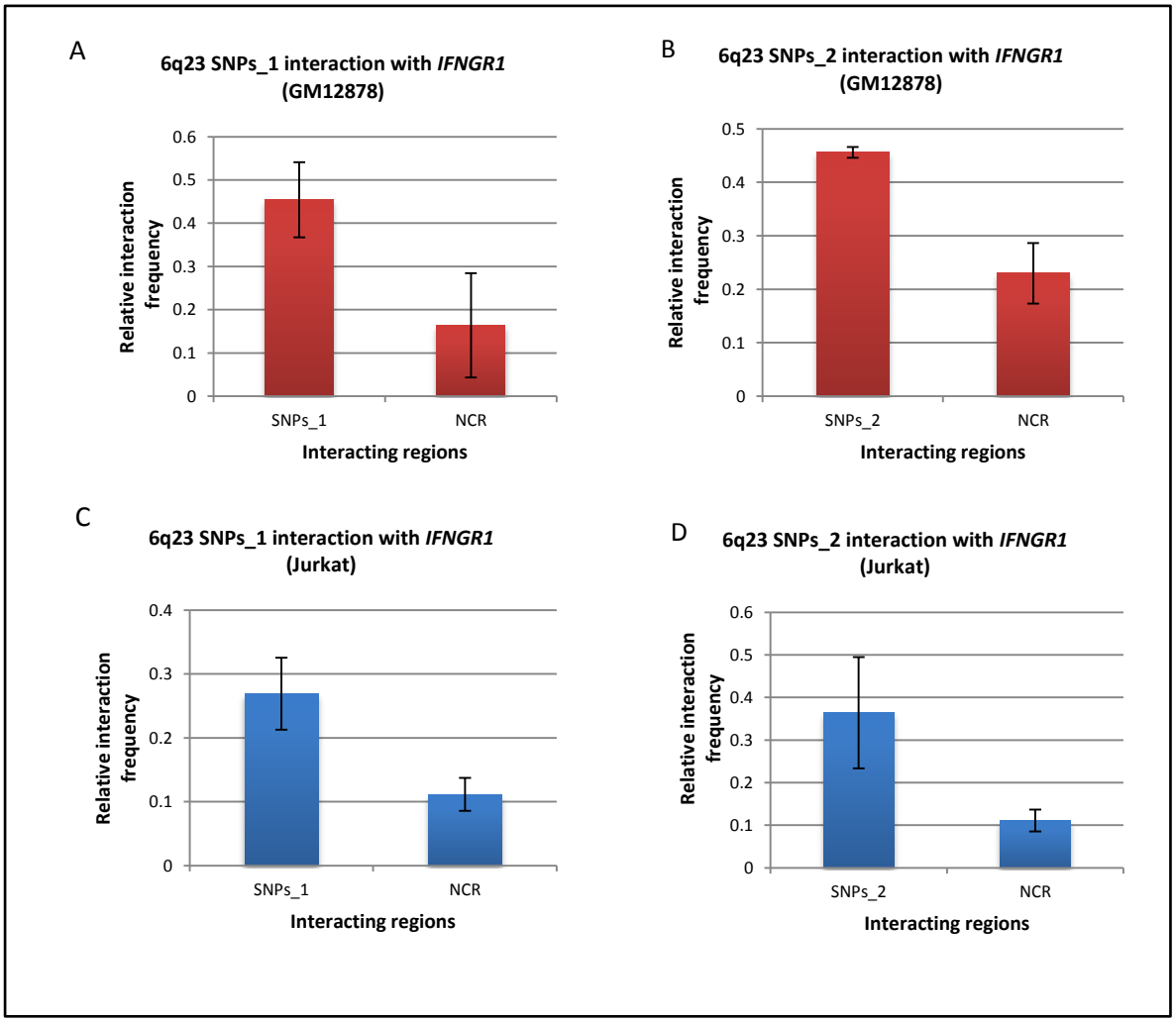
2. Upstream *IFNGR1* region interacts with 6q23 SNPs in Jurkat T-cells, primary human synovial fibroblasts and in a genotype-specific manner in B-cells

Interactions between a *Hin*dIII fragment located upstream of *IFNGR1* (IFNGR1_1 chr6:137570290-137583223) and *Hin*dIII fragments located in an LD block containing RA associated SNPs (SNPs_1 chr6:137952897-137959707, SNPs_2 chr6:137959709-137963083). These interactions were present in both cell lines, with a 1% FDR, and were only detected in the region capture experiment since the *IFNGR1 Hin*dIII fragment maps to a non-coding region upstream of *IFNGR1* so no promoters were involved.

The SNPs_1 and SNPs_2 fragments are located at the 5' region of the RA LD block and showed only a low level of interaction in the intial assays. The relative interaction frequency between *IFNGR1* and the SNPs_1 region was significantly increased over the NCR in GM12878 cells (p=0.011) (Figure 45A) but not in Jurkats (p=0.075) (Figure 45C), however, there was an increase in interaction with the test region compared to the NCR. In both cell lines the relative interaction frequency between *IFNGR1* and the SNPs_2 region was not significantly different to the NCR - GM12878 (p=0.062) (Figure 45B) and Jurkats (p=0.085) (Figure 45D). However, there was an increase in interaction with the test region compared to the NCR in both cell lines.

3C-qPCR using the downstream SNP fragment primers showed that the interaction peak localised to the *Hin*dIII fragment containing the rs6927172 SNP (Figure 46). Assays were performed using 3C libraries generated from HapMap B-LCLs specific for the appropriate SNP genotype - GM12878, GM12145 (CG); GM11993, GM10838, GM12892 (CC); GM07037, GM10850, GM10858 (GG) (Figure 46A). The interaction occurred more frequently in samples containing the risk allele (G) compared to the samples that were homozygous for the non-risk (C) allele (p=0.04) (Figure 46B). Jurkat T-cells (Figure 46C) and primary human synovial fibroblasts (Figure 46D) were also analysed, with the interaction peak localising to the same fragment as the B-LCLs.

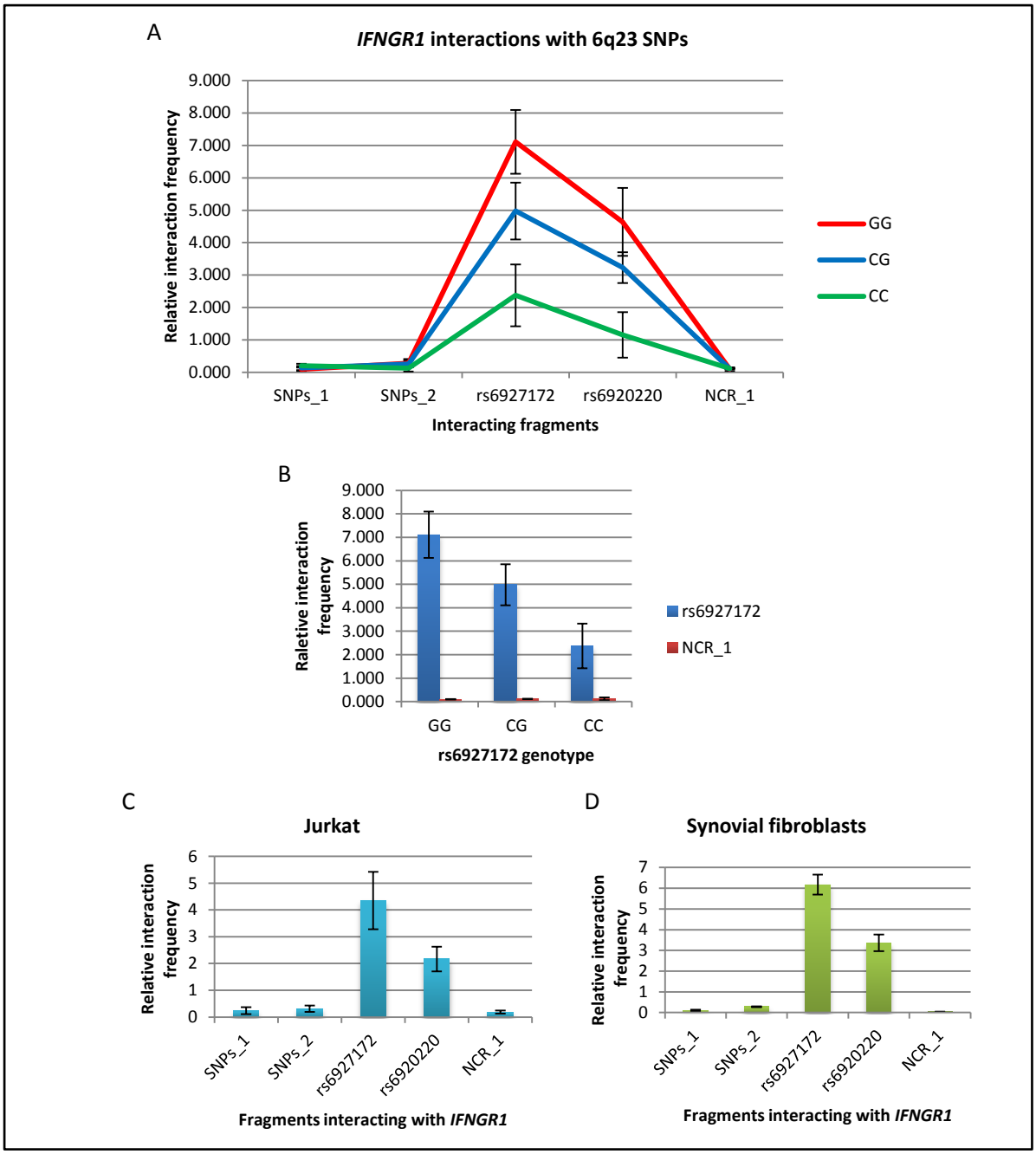
Figure 45: Initial 3C-qPCR analysis of *IFNGR1* interactions with the 6q23 SNP region



3C-qPCR was carried out using the anchor primer IFNGR1_1 in combination with a test primer (SNPs_1 or SNPs_2) designed for the interacting *Hin*dIII fragment or a non-interacting region (NCR). For each set of primers a standard curve was generated using BAC control libraries spanning the region of interest. SYBR green qPCR was carried out in triplicate using 50ng of 3C library per reaction or a no-template (water) control. Triplicate 3C libraries were analysed and T-tests performed to determine if the relative interaction frequency in the test fragment was significantly different to the NCR.

(A) Interaction between *IFNGR1* and a *Hin*dIII site located in the RA SNPs LD block (SNPs_1) was analysed in GM12878 cells; (B) Interaction between *IFNGR1* and a *Hin*dIII site located in the RA SNPs LD block (SNPs_2) was analysed in GM12878 cells; (C) Interaction between *IFNGR1* and a *Hin*dIII site located in the RA SNPs LD block (SNPs_1) was analysed in Jurkats; (B) Interaction between *IFNGR1* and a *Hin*dIII site located in the RA SNPs LD block (SNPs_2) was analysed in Jurkats.

Figure 46: Interactions between *IFNGR1* and the 6q23 SNP region



3C-qPCR was carried out using the anchor primer IFNGR1_1 in combination with primers designed in multiple *Hin*dIII fragments within the RA SNPs LD block - SNPs_1, SNPs_2, a *Hin*dIII fragment containing the rs6920220 RA associated SNP, a *Hin*dIII fragment containing the putative functional SNP rs6927172, along with a NCR. For each set of primers (anchor fragment primer + test region or NCR) a standard curve was generated using BAC control libraries spanning the region of interest. SYBR green qPCR was carried out in triplicate using 50ng of 3C library per reaction or a no-template control.

Assays were performed using 3C libraries generated from HapMap B-LCLs specific for the appropriate rs6927172 SNP genotype - GM12878, GM12145 (CG); GM11993, GM10838, GM12892 (CC); GM07037, GM10850, GM10858 (GG). T-tests were used to compare samples homozygous for the risk allele (CC) and samples homozygous for the non-risk (GG) allele (p=0.040): (A) Genotype-specific interactions in B-LCLs at the different *Hin*dIII fragments, (B) Genotype-specific interactions at the *Hin*dIII fragment containing the rs6927172 SNP. Jurkat T-cells (C) and primary human synovial fibroblasts (D) were also analysed. The data represents the average interaction frequencies of the samples tested; error bars are +/- St.Dev.

3. 6q23 SNPs interact with lncRNAs downstream of *TNFAIP3* in B-LCLs, T-cells and primary human synovial fibroblasts

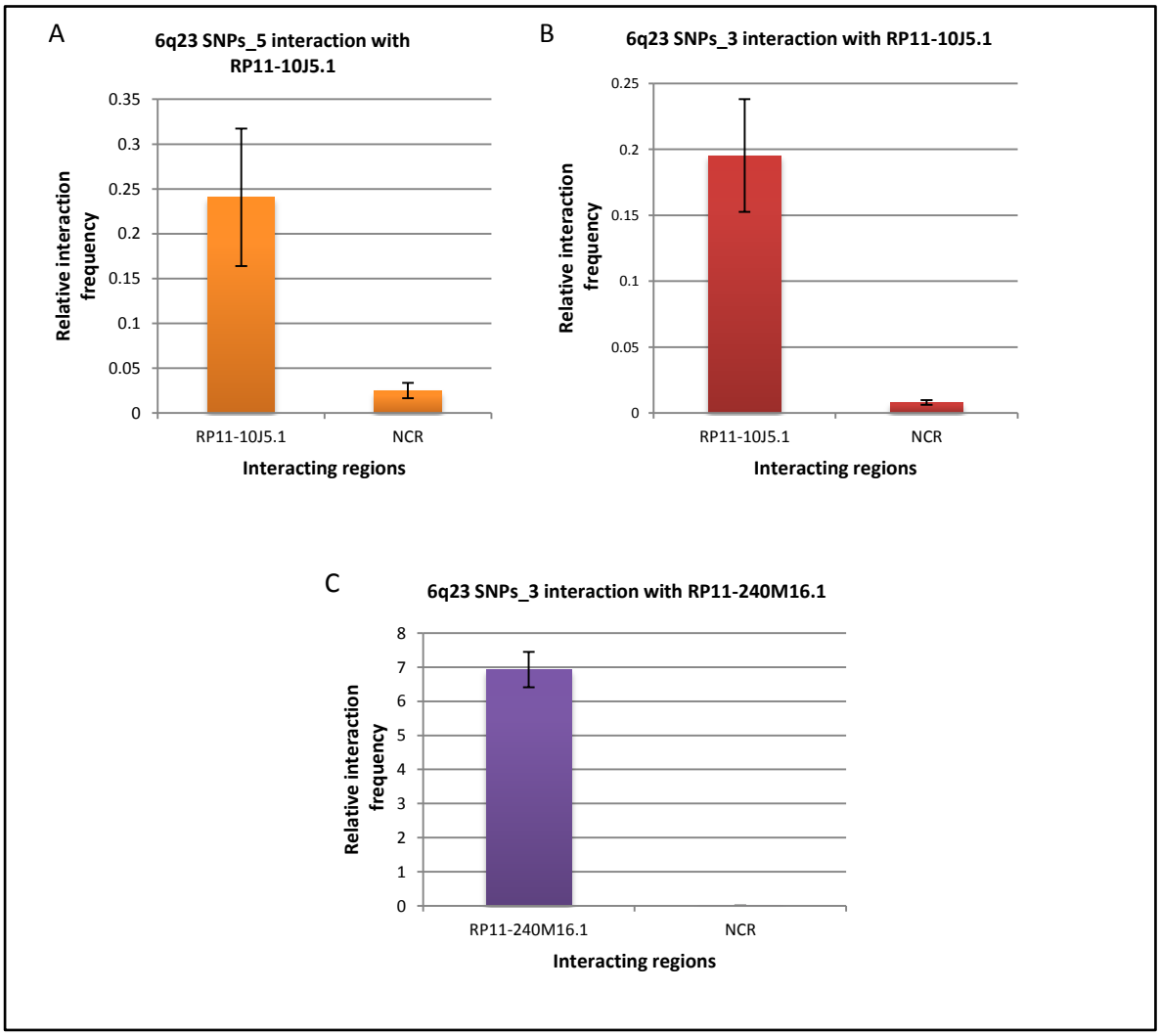
Interactions were detected between *Hin*dIII fragments located downstream from *TNFAIP3* at the lncRNAs RP11-10J5.1 (lncRNA_1 chr6:138262495-138267565) and RP11-240M16.1 (lncRNA_3 chr6:138267567-138268650) and a *Hin*dIII fragment located in the LD block containing RA associated SNPs (SNPs_3 chr6:137983020-137989382). These interactions were observed in only the GM12878 cells, in the region capture experiment with a 5% FDR whereas the interactions with *IL20RA* and *IFNGR1* were detected with the more stringent FDR of 1%. In the promoter capture experiment, with an FDR of 1%, an interaction was detected between a *Hin*dIII fragment located downstream from *TNFAIP3* at the lncRNA RP11-10J5.1 (lncRNA_1 chr6:138262495-138267565) and a *Hin*dIII fragment located in the LD block containing RA associated SNPs (SNPs_5 chr6:138007203-138017056).

Initially, 3C-qPCR was performed on triplicate 3C libraries generated from GM12878 B-LCLs only as the interaction was not observed in T-cells in the Capture Hi-C experiment. T-tests were performed to determine statistical significance ($p<0.05$) which showed that the relative interaction frequencies between the SNP region and the lncRNAs were significantly increased over the NCR, confirming these interactions (Figure 47A-C).

Following on from the initial analysis, further assays were performed to assess the interactions between the SNPs LD block fragments and the RP11-10J5.1 and RP11-240M16.1 lncRNAs. 3C-qPCR using the SNP fragment primers showed that the RP11-10J5.1 lncRNA_1 interaction could not be localised to one particular fragment, instead the interaction was located between the rs6927172 SNP *Hin*dIII fragment and the SNPs_5 fragment 10kb downstream of this fragment (Figure 48A). In contrast, the RP11-240M16.1 lncRNA_3 interaction peak clearly localised to the rs6920220 RA SNP *Hin*dIII fragment (Figure 49A).

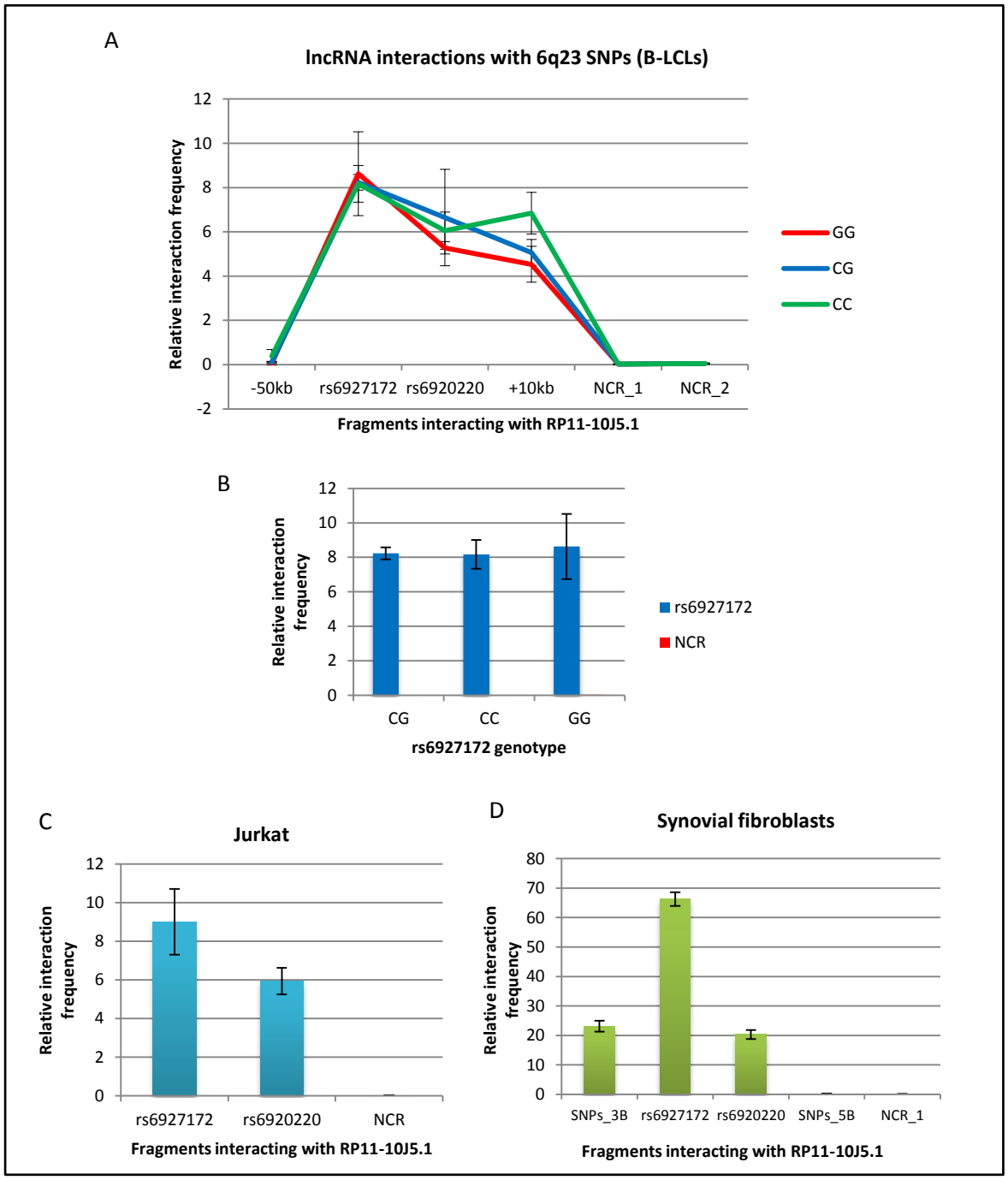
Assays were subsequently performed using 3C libraries generated from HapMap B-LCLs specific for the appropriate rs6927172 SNP genotype - GM12878, GM12145, GM11994 (CG); GM11993, GM10838, GM12892 (CC); GM07037, GM10850, GM10858 (GG) (Figure 48A-B and Figure 49A). No significant difference in interaction with the RP11-10J5.1 or the RP11-240M16.1 fragments was observed between B-LCLs with the different rs6927172 genotypes. Analysis of primary human synovial fibroblasts also demonstrated interactions with the lncRNA fragments (Figure 48D and Figure 49C). Jurkat cells were also analysed and showed clear interactions despite not being detected in the initial analysis of the Capture Hi-C data (Figure 48C and Figure 49B).

Figure 47: Initial 3C-qPCR analysis of downstream *TNFAIP3* lncRNA interactions with the 6q23 SNP region



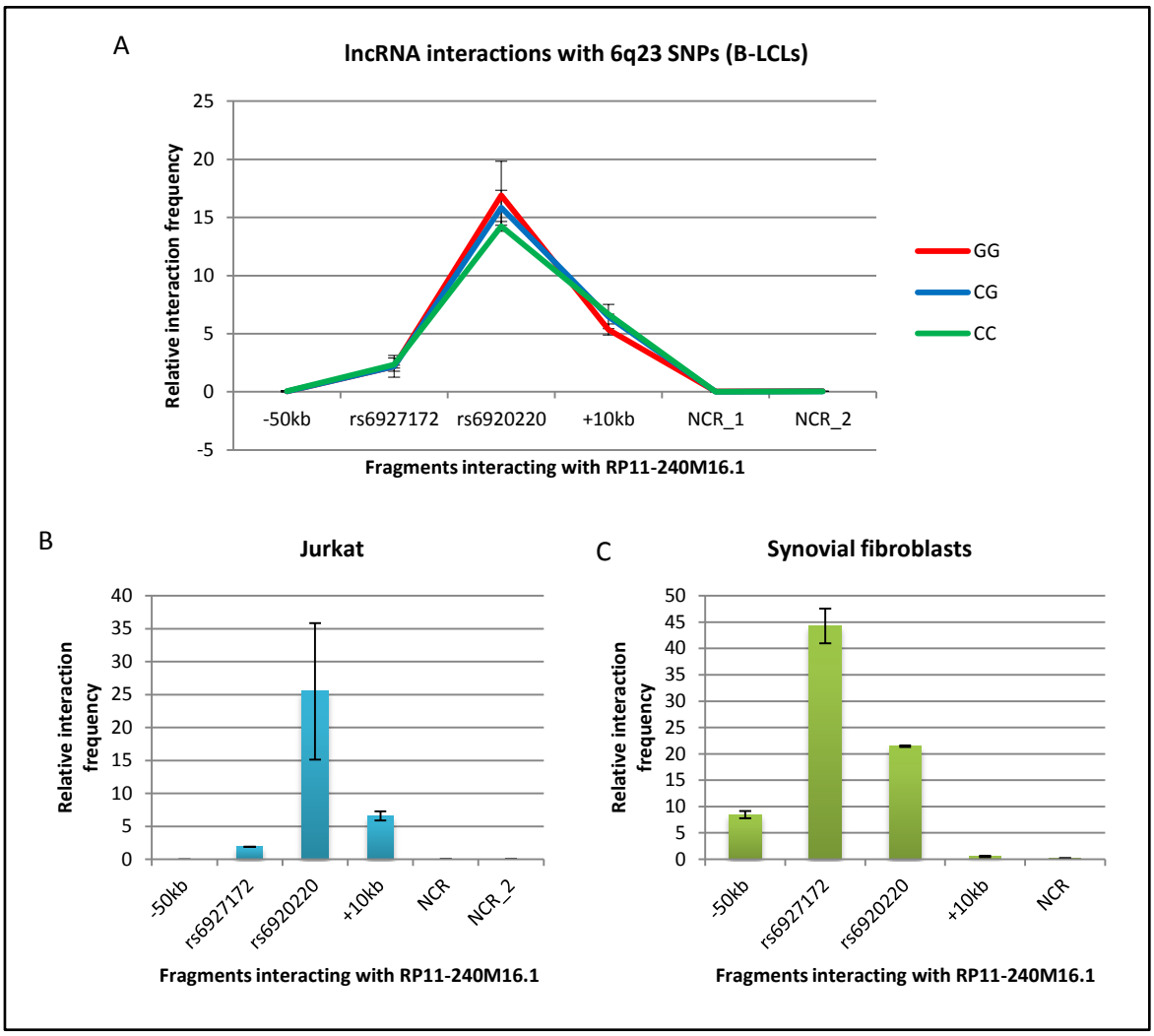
3C-qPCR was carried out using the anchor primer SNPs_3 or SNPs_5B in combination with the test primers lncRNA_1B (representing RP11-10J5.1), and lncRNA_3 (representing RP11-240M16.1) or a non-interacting region (NCR). For each set of primers a standard curve was generated using BAC control libraries spanning the region of interest. SYBR green qPCR was carried out in triplicate using 50ng of 3C library per reaction or a no-template (water) control. Triplicate 3C libraries were analysed and T-tests performed to determine if the relative interaction frequency in the test fragment was significantly different to the NCR.
(A) Interaction between a *Hin*dIII site in the RP11-10J5.1 lncRNA (lncRNA_1B) and a *Hin*dIII site located in the RA SNPs LD block (SNPs_3); (B) Interaction between a *Hin*dIII site in the RP11-10J5.1 lncRNA (lncRNA_1B) and a *Hin*dIII site located in the RA SNPs LD block (SNPs_5B); (C) Interaction between a *Hin*dIII fragment in the middle of the RP11-240M16.1 lncRNA (lncRNA_3) and a *Hin*dIII site located in the RA SNPs LD block (SNPs_3).

Figure 48: Interactions between lncRNA RP11-10J5.1 and the 6q23 SNP region



3C-qPCR was carried out using the anchor primer lncRNA_1B in combination with primers designed in multiple *Hin*dIII fragments within the RA SNPs LD block - SNPs_3B, SNPs_5B, a *Hin*dIII fragment containing the rs6920220 RA associated SNP, a *Hin*dIII fragment containing the putative functional SNP rs6927172, along with NCRs. For each set of primers a standard curve was generated using BAC control libraries spanning the region of interest. SYBR green qPCR was carried out in triplicate using 50ng of 3C library per reaction or a no-template control.
Assays were performed using 3C libraries generated from HapMap B-LCLs with the appropriate rs6927172 SNP genotype - GM12878, GM12145, GM11994 (CG); GM11993, GM10838, GM12892 (CC); GM07037, GM10850, GM10858 (GG). (A) Genotype-specific interactions in B-LCLs at different *Hin*dIII fragments, (B) Genotype-specific interactions at the *Hin*dIII fragment containing rs6927172, (C) Jurkats, (D) synovial fibroblasts (only two libraries available). The data represents the average interaction frequencies of the samples tested; error bars are +/- St.Dev.

Figure 49: Interactions between lncRNA RP11-240M16.1 and the 6q23 SNP region



3C-qPCR was carried out using the anchor primer lncRNA_3 in combination with primers designed in multiple *Hin*dIII fragments within the RA SNPs LD block - SNPs_3B, SNPs_5B, a *Hin*dIII fragment containing the rs6920220 RA associated SNP, a *Hin*dIII fragment containing the putative functional SNP rs6927172, along a NCR. For each set of primers a standard curve was generated using BAC control libraries spanning the region of interest. SYBR green qPCR was carried out in triplicate using 50ng of 3C library per reaction or a no-template control.
Assays were performed using 3C libraries generated from HapMap B-LCLs specific for the appropriate rs6927172 SNP genotype - GM12878, GM12145, GM11994 (CG); GM11993, GM10838, GM12892 (CC); GM07037, GM10850, GM10858 (GG). (A) Genotype-specific interactions in B-LCLs at the different *Hin*dIII fragments, (B) Interactions in Jurkats, (C) Interactions in primary human synovial fibroblasts (only two libraries available). The data represents the average interaction frequencies of the samples tested; error bars are +/- St.Dev.

4. 6q23 SNPs interact with a region downstream of lncRNAs in B-LCLs, T-cells and primary human synovial fibroblasts

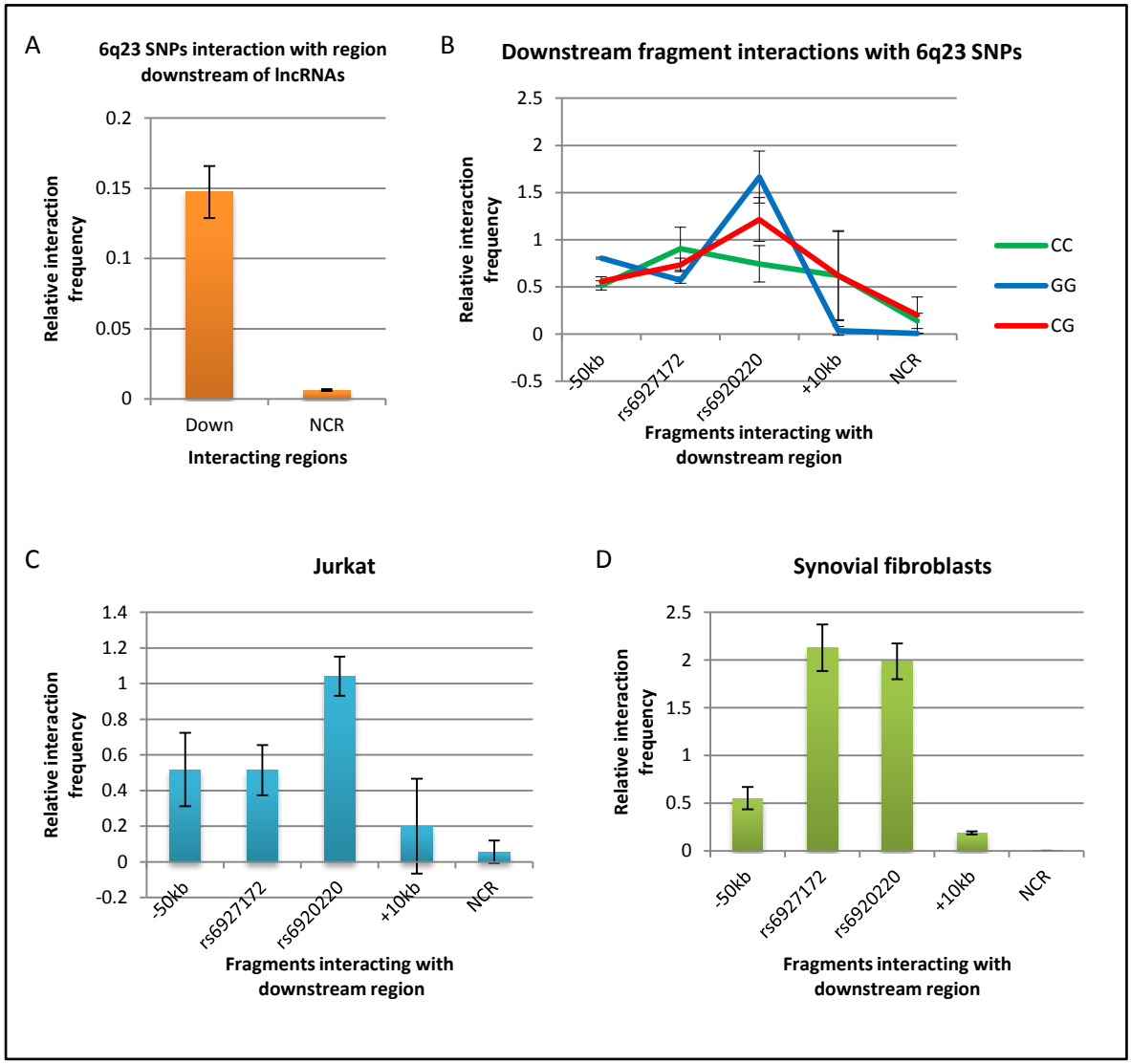
Interactions were detected between a *Hin*dIII fragment located downstream of the *TNFAIP3* lncRNAs (DOWN chr6:138320836-138334122) and a *Hin*dIII fragment located in the LD block containing RA associated SNPs (SNPs_3 chr6:137983020-137989382). This interaction was observed in only the GM12878 cells, in the region capture experiment only and with a 5% FDR whereas the interactions with *IL20RA* and *IFNGR1* were detected with the more stringent FDR of 1%.

3C-qPCR was performed on triplicate 3C libraries generated from GM12878 B-LCLs only as the interaction was not observed in T-cells in the Capture Hi-C experiment (Figure 50A). T-tests were performed to determine statistical significance ($p<0.05$) which showed that the relative interaction frequencies between the SNP region and the downstream fragment was significantly increased over the NCR ($p=0.007$), confirming this interaction.

Following on from the initial analysis, further assays were performed to assess the interactions between the SNPs LD block fragments and the downstream *Hin*dIII *fragment*. 3C-qPCR using the SNP fragment primers showed that the interaction interaction could not be localised to one particular fragment, instead the interaction was located between the rs6927172 SNP *Hin*dIII fragment, rs6920220 SNP *Hin*dIII fragment and the SNPs_5 fragment 10kb downstream of this fragment (Figure 49B).

Assays were performed using 3C libraries generated from HapMap B-LCLs specific for the appropriate SNP genotype - GM12878, GM12145, GM11994 (CG); GM11993, GM10838, GM12892 (CC); GM07037, GM10850, GM10858 (GG). No significant difference in interaction with the downstream fragment was observed between B-LCLs with the different rs6927172 genotypes (Figure 50B). Jurkat cells were also analysed and showed clear interactions despite the interaction with this cell line not being detected in the initial analysis of the Capture Hi-C data (Figure 50C). Primary human synovial fibroblasts also showed interactions with the lncRNAs (Figure 50D).

Figure 50: 3C-qPCR analysis of downstream *TNFAIP3* interactions with the 6q23 SNP region



3C-qPCR was carried out using the DOWN anchor primer in combination with the test primer SNPs_3 or a non-interacting region (NCR) (A). For each set of primers a standard curve was generated using BAC control libraries spanning the region of interest. SYBR green qPCR was carried out in triplicate using 50ng of 3C library per reaction or a no-template (water) control. Triplicate 3C libraries were analysed and T-tests were performed to determine if the relative interaction frequency of the test fragment was significantly different to the NCR, shown in the table ($p < 0.05$).
Further 3C-qPCR was carried out using the DOWN anchor primer in combination with primers designed in multiple *Hin*dIII fragments within the RA SNPs LD block - SNPs_3B, SNPs_5B, a *Hin*dIII fragment containing the rs6920220 RA associated SNP, a *Hin*dIII fragment containing the putative functional SNP rs6927172, along with a NCR (B). Assays were performed using 3C libraries from HapMap B-cell lines specific for the appropriate rs6927172 SNP genotype - GM11994, GM10831, GM06993 (CG), GM12892, GM12707, GM10838 (CC), GM07037, GM10850, GM10858 (GG). Jurkats (C) and primary human synovial fibroblasts (D) were also analysed. The data represents the average interaction frequencies of the samples tested; error bars are +/- St.Dev.

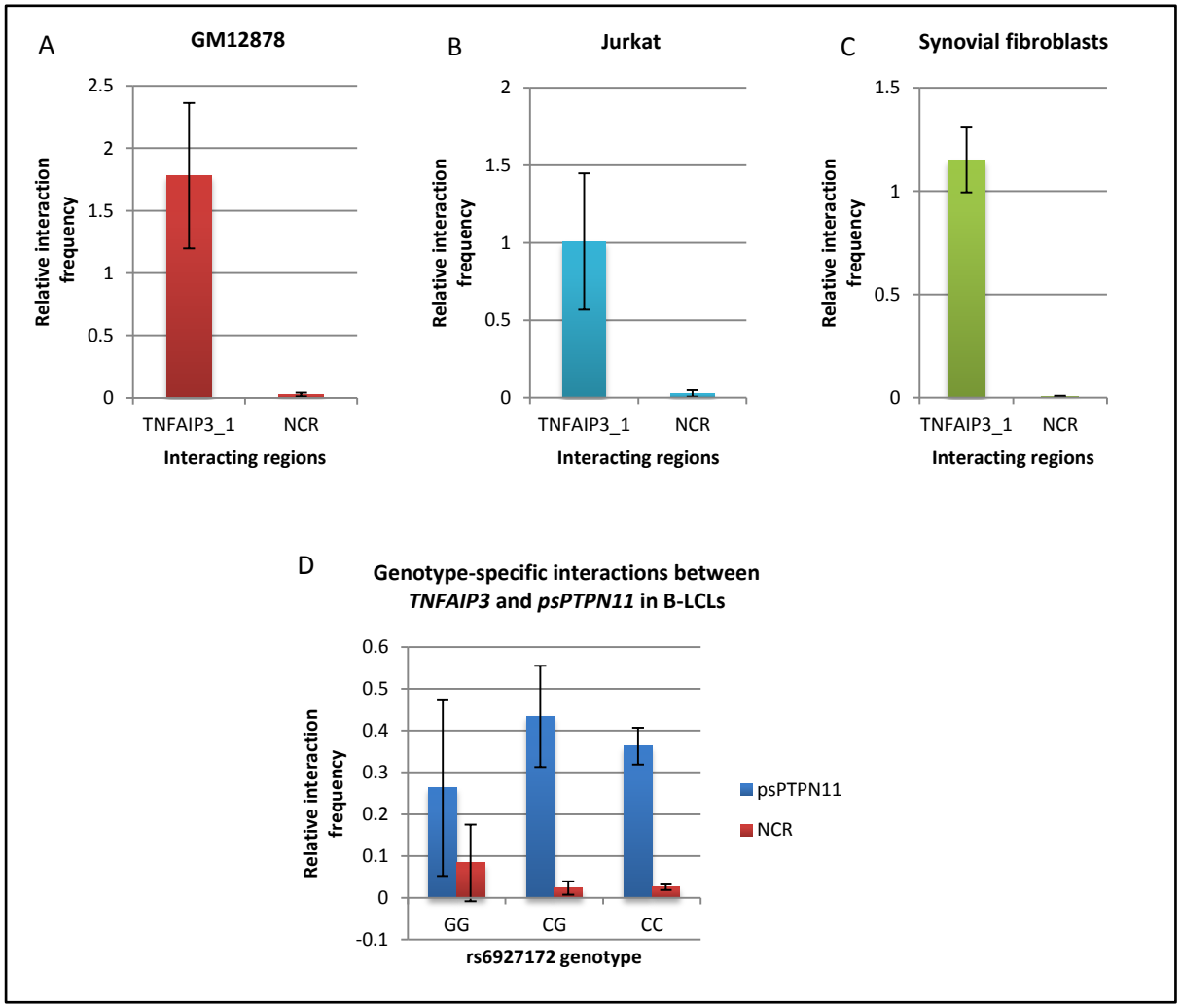
4.1.2. Non-SNP interesting interactions

1. *TNFAIP3* interacts with a *PTPN11* pseudogene (*psPTPN11*) in GM12878, Jurkats and synovial fibroblasts

An interaction was observed between a *Hin*dIII fragment 3' of the SNPs LD block containing a *PTPN11* pseudogene (psPTPN11_1 chr6:138025956-138036419) and a *Hin*dIII fragment located within *TNFAIP3* (TNFAIP3_1 chr6:138192730-138193357). This interaction was observed in GM12878 B-LCLs and Jurkat T-cells, in the region capture experiment only and with a 1% FDR.

3C-qPCR was performed on triplicate samples for GM12878 and Jurkats (Figure 51A and B), and T-tests performed to determine statistical significance ($p<0.05$). This interaction was also detected in synovial fibroblasts, however only one sample was available for analysis (Figure 51C). Interaction between 3' LD block/*psPTPN11* and *TNFAIP3* was significantly increased over the NCR in GM12878 ($p=0.03$), Jurkat cells ($p=0.05$) and was increased relative to the NCR in synovial fibroblasts. Assays were also performed using 3C libraries generated from HapMap B-LCLs specific for the appropriate rs6927172 SNP genotype - GM12878, GM12145, GM11994 (CG); GM11993, GM10838, GM12892 (CC); GM07037, GM10850, GM10858 (GG). No significant difference in interaction with *psPTPN11* was observed between B-LCLs with the different rs6927172 genotypes (Figure 51D).

Figure 51: 3C-qPCR analysis of *TNFAIP3* interactions with the *PTPN11* pseudogene



For each set of primers a standard curve was generated using BAC control libraries spanning the region of interest. SYBR green qPCR was carried out in triplicate using 50ng of 3C library per reaction or a no-template (water) control. Triplicate 3C libraries were analysed for GM12878 (A) and Jurkats (B) and a single synovial fibroblast library (C) (3 technical replicates to generate error bars). T-tests were performed to determine if the relative interaction frequency of the test fragment was significantly different to the NCR. Assays were also performed using 3C libraries from HapMap B-cell lines specific for the appropriate rs6927172 SNP genotype (D) - GM11994, GM10831, GM06993 (CG), GM12892, GM12707, GM10838 (CC), GM07037, GM10850, GM10858 (GG). The data represents the average interaction frequencies of the samples tested; error bars are +/- St.Dev.

2. *TNFAIP3* interacts with lncRNAs RP11-10J5.1 and RP11-240M16.1 in GM12878, Jurkats and synovial fibroblasts

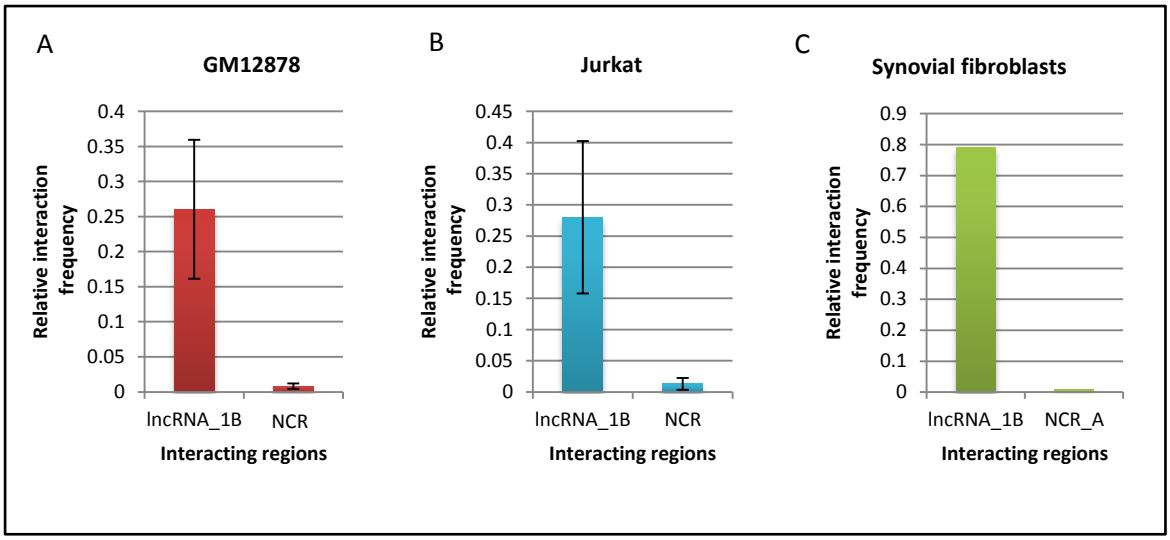
An interaction was detected between a *Hin*dIII fragment located within *TNFAIP3* (TNFAIP3_1 chr6:138192730-138193357) and a *Hin*dIII fragment containing the RP11-10J5.1 lncRNA (lncRNA_1 chr6:138262495-138267565). Interactions were also detected between a *Hin*dIII fragment located at the 3' end of *TNFAIP3* (TNFAIP3_3 chr6:138202662-138204711) and *Hin*dIII fragments containing the lncRNAs RP11-10J5.1 lncRNA (lncRNA_1 chr6:138262495-138267565) and RP11-240M16.1 (lncRNA_3 chr6:138267567-138268650). These interactions were observed in GM12878 B-LCLs and Jurkat T-cells, in the region capture experiment only and with a 1% FDR.

3C-qPCR was performed on triplicate samples for each cell line and T-tests performed to determine statistical significance ($p<0.05$). The interaction between TNFAIP3_1 and lncRNA RP11-10J5.1 was significantly increased over the NCR in GM12878 cells ($p=0.048$) (Figure 52A). The interaction in Jurkat cells was not significantly different to the NCR interaction ($p=0.06$) (Figure 52B), however the interaction frequency was higher in the test region compared to the non-interacting region. This interaction was also detected in synovial fibroblasts, however only one sample was available for analysis (Figure 52C).

The relative interaction frequency of the interaction between TNFAIP3_3 and both RP11-10J5.1 and RP11-240M16.1 was not significantly different to the NCR in GM12878 cells, however the interaction frequency in the test region was higher than the non-interacting region (Figure 53A and D). The interaction between TNFAIP3_3 and both RP11-10J5.1 and RP11-240M16.1 in Jurkat cells was significantly higher than the NCR ($p=0.006$ and $p=0.01$ respectively) (Figure 53B and E). The interactions between TNFAIP3_3 and both RP11-10J5.1 and RP11-240M16.1 were also detected in synovial fibroblasts (Figure 53C and F).

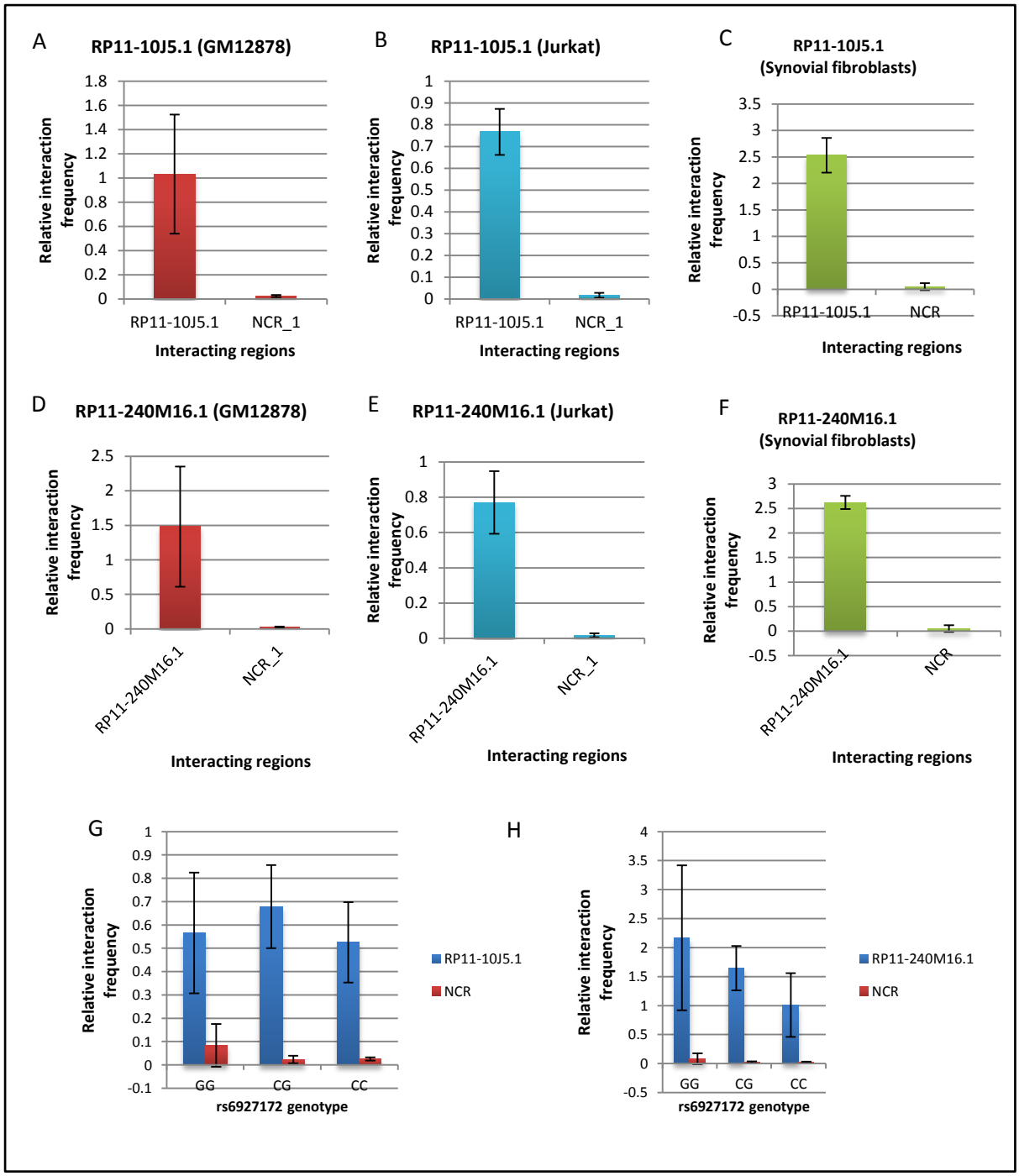
Assays were also performed using 3C libraries generated from HapMap B-LCLs specific for the appropriate SNP genotype (Figure 53G and H) - GM12878, GM12145, GM11994 (CG); GM11993, GM10838, GM12892 (CC); GM07037, GM10850, GM10858 (GG). No significant difference in interaction with either lncRNA fragment was observed between B-LCLs with the different rs6927172 genotypes.

Figure 52: 3C-qPCR analysis of *TNFAIP3* interactions with RP11-10J5.1



3C-qPCR was carried out using the TNFAIP3_1 anchor primer with the lncRNA_1B test primer and a NCR primer. For each set of primers a standard curve was generated using BAC control libraries spanning the region of interest. SYBR green qPCR was carried out in triplicate using 50ng of 3C library per reaction or a no-template (water) control. Triplicate 3C libraries were analysed for GM12878 (A) and Jurkats (B) and T-tests performed to determine if the relative interaction frequency of the test fragment was significantly different to the NCR. The data represents the average interaction frequencies of the samples tested; error bars are +/- St.Dev. Only one library was analysed for the synovial fibroblasts (C).

Figure 53: 3C-qPCR analysis of *TNFAIP3* interactions with RP11-10J5.1 and RP11-240M16.1



3C-qPCR was carried out using the TNFAIP3_3 anchor primer with the lncRNA_1B or lncRNA_3B test primers and NCR primers. For each set of primers a standard curve was generated using BAC control libraries spanning the region of interest. SYBR green qPCR was carried out in triplicate using 50ng of 3C library per reaction or a no-template (water) control. Triplicate 3C libraries were analysed for GM12878 (A and D) and Jurkats (B and E) and T-tests performed to determine if the relative interaction frequency of the test fragment was significantly different to the NCR. The data represents the average interaction frequencies of the samples tested; error bars are +/- St.Dev. Only one library was analysed for the synovial fibroblasts (C and F). Assays were also performed using 3C libraries from HapMap B-cell lines specific for the appropriate rs6927172 SNP genotype (G-H) - GM11994, GM10831, GM06993 (CG), GM12892, GM12707, GM10838 (CC), GM07037, GM10850, GM10858 (GG).

3. *TNFAIP3* interacts with a Y_RNA in GM12878 B-LCLs in the promoter capture experiment

An interaction was observed between a *Hin*dIII fragment containing a Y-RNA (chr6:138105291-138121041) and a *Hin*dIII fragment located 5' of *TNFAIP3* (chr6:138184709-138186854). These interactions were observed in GM12878 B-LCLs in the promoter capture experiment with a 1% FDR. 3C-qPCR was performed on triplicate samples for GM12878 B-LCLs and duplicate samples for Jurkat T-cells and T-tests performed to determine statistical significance (p<0.05). Interaction between *TNFAIP3* and Y_RNA was significantly increased over NCR in GM12878 cells (p=0.008) (Figure 54A), but not in Jurkats (p=0.228) (Figure 54B) although the interaction frequency in the test region was higher than the non-interacting region (only two biological replicates available for Jurkats). No significant difference in interaction with either lncRNA fragment was observed between B-LCLs with the different rs6927172 genotypes (Figure 54C). Synovial fibroblasts were unavailable for this assay.

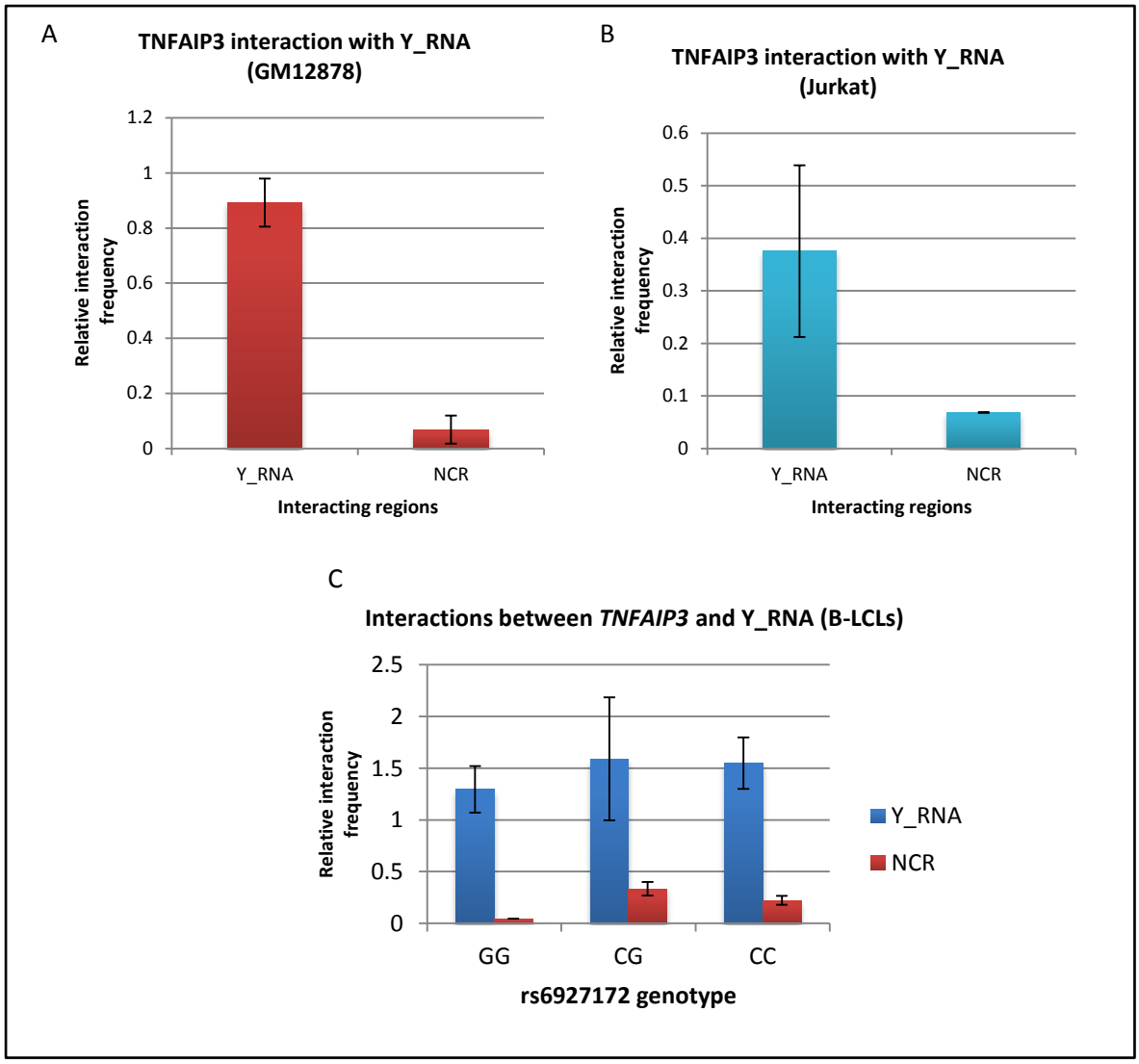
4. *IL20RA* interacts with *TNFAIP3* in GM12878 B-LCLs in the promoter capture experiment

An interaction was detected between a *Hin*dIII fragment containing the *IL20RA* promoter (IL20RA_2 chr6:137421229-137423210) and a *Hin*dIII fragment located at the 5' end of *TNFAIP3* (TNFAIP3_2 chr6:138186856-138192635). These interactions were observed in GM12878 B-LCLs in the promoter capture experiment only and with a 1% FDR, however interactions in Jurkats were also analysed in the follow-up 3C-qPCR experiments. 3C-qPCR was performed on triplicate samples for each cell line and T-tests were performed to determine statistical significance (p<0.05). The interaction between IL20RA_3 and TNFAIP3_2 was not significantly increased over NCR in GM12878 (p=0.084) (Figure 55A) and Jurkat T-cells (p=0.216) (Figure 55B), however the interaction frequency was higher than in the non-interacting region. The relative interaction frequency for this interaction was low in both cell lines and was not detected in synovial fibroblasts (Figure 55C). No significant difference in interaction frequency was observed between B-LCLs with the different rs6927172 genotypes (Figure 55D).

5. *IL20RA* interacts with lncRNA RP11-10J5.1 in GM12878 B-LCLs

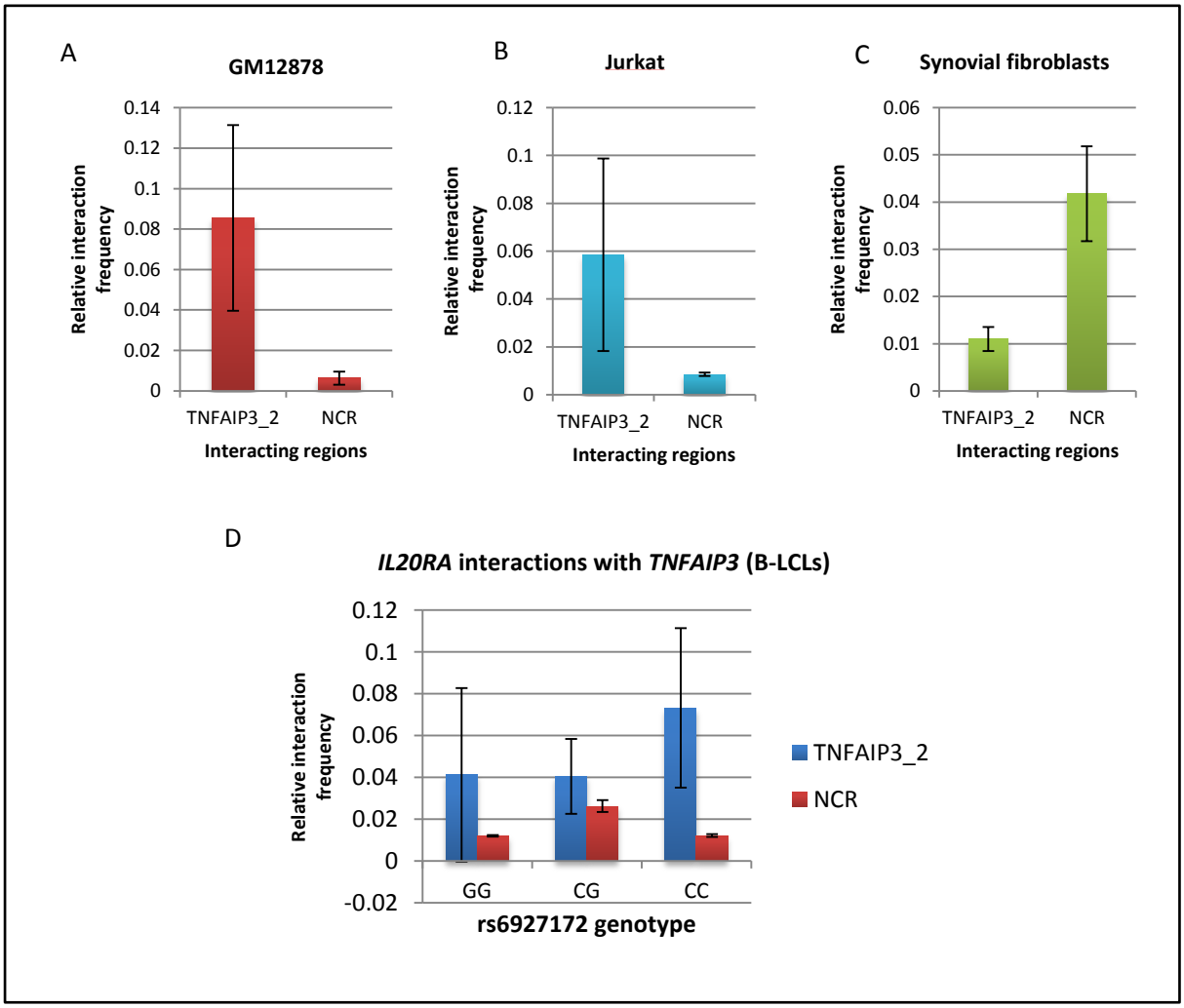
An interaction was detected between a *Hin*dIII fragment containing the *IL20RA* promoter (chr6:137421229-137423210) and a *Hin*dIII fragment located in the downstream *TNFAIP3* lncRNA RP11-10J5.1 (lncRNA_1 chr6:138262495-138267565). This interaction was observed in GM12878 B-LCLs in the promoter capture experiment only and with a 1% FDR, however interactions in Jurkats were also analysed in the follow-up 3C-qPCR experiments. 3C-qPCR was performed on triplicate samples for each cell line and T-Tests performed to determine statistical significance (p < 0.05). The interaction between *IL20RA* and lncRNA RP11-10J5.1 was not significantly different to the NCR in GM12878 (p=0.062) (Figure 56A) and Jurkat T-cells (p=0.066) (Figure 56B), although the interaction frequency was higher than in the non-interacting region. This interaction was also detected in synovial fibroblasts (Figure 56C). No significant difference in interaction frequency was observed between B-LCLs with the different rs6927172 genotypes (Figure 56D).

Figure 54: 3C-qPCR analysis of *TNFAIP3* interactions with *Y_RNA*



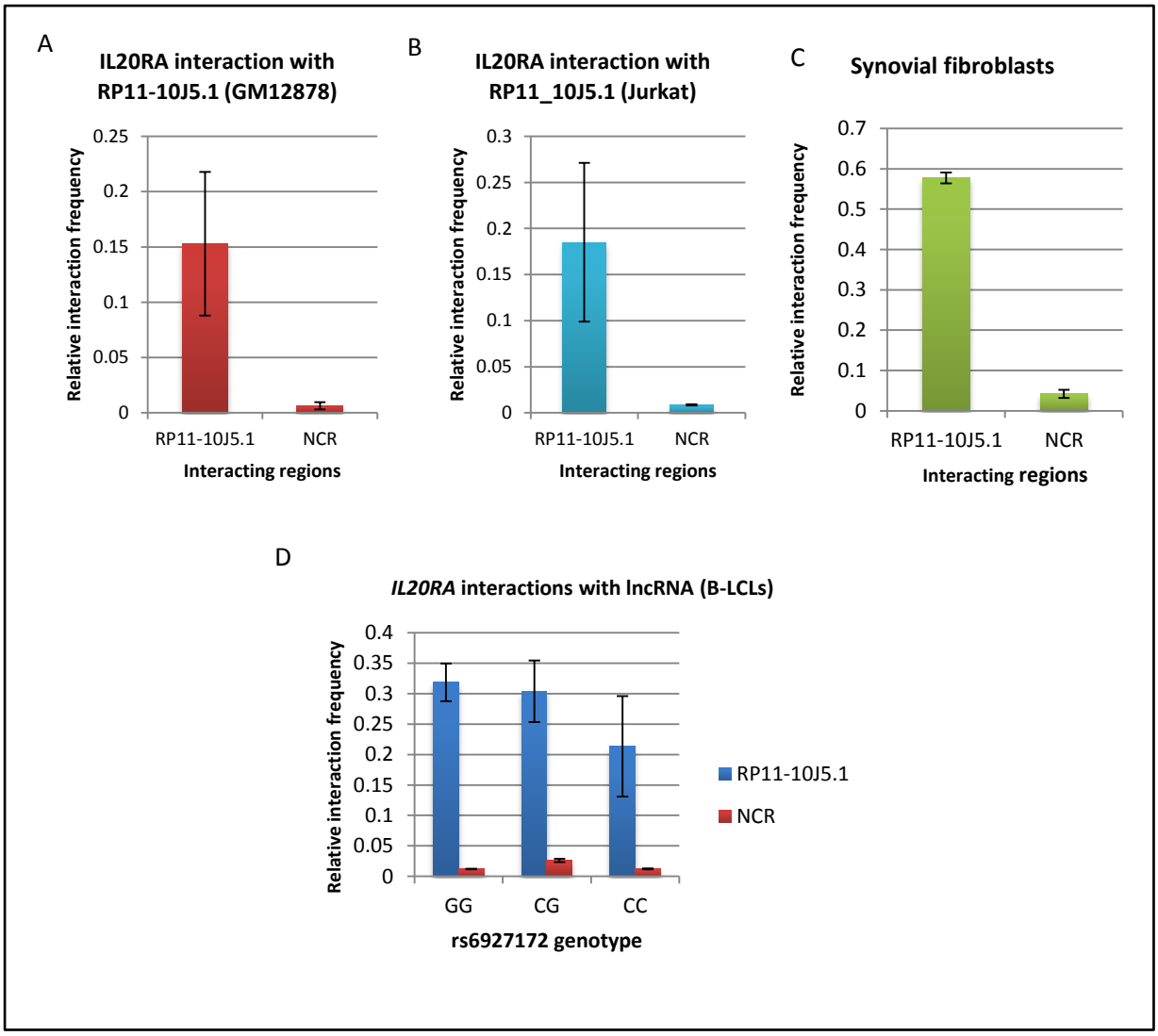
3C-qPCR was carried out using the TNFAIP3_4 anchor primer with the Y_RNA test primers and NCR primers. For each set of primers a standard curve was generated using BAC control libraries spanning the region of interest. SYBR green qPCR was carried out in triplicate using 50ng of 3C library per reaction or a no-template (water) control. Triplicate 3C libraries were analysed for GM12878 (A) and duplicate libraries for Jurkats (B) and T-tests performed to determine if the relative interaction frequency of the test fragment was significantly different to the NCR. The data represents the average interaction frequencies of the samples tested; error bars are +/- St.Dev. Assays were also performed using 3C libraries from HapMap B-cell lines specific for the appropriate rs6927172 SNP genotype (C) - GM11994, GM10831, GM06993 (CG), GM12892, GM12707, GM10838 (CC), GM07037, GM10850, GM10858 (GG). No significant differences were observed between the genotypes.

Figure 55: 3C-qPCR analysis of *IL20RA* interactions with *TNFAIP3*



3C-qPCR was carried out using the TNFAIP3_2 anchor primer with the IL20RA_3 test primers and NCR primers. For each set of primers a standard curve was generated using BAC control libraries spanning the region of interest. SYBR green qPCR was carried out in triplicate using 50ng of 3C library per reaction or a no-template (water) control. Triplicate 3C libraries were analysed for GM12878 (A) and Jurkats (B) and T-tests were performed to determine if the relative interaction frequency of the test fragment was significantly different to the NCR. The data represents the average interaction frequencies of the samples tested; error bars are +/- St.Dev. Only one library was analysed for the synovial fibroblasts (C) (error bars from technical replicates). Assays were also performed using 3C libraries from HapMap B-cell lines specific for the appropriate rs6927172 SNP genotype (D) - GM11994, GM10831, GM06993 (CG), GM12892, GM12707, GM10838 (CC), GM07037, GM10850, GM10858 (GG).

Figure 56: 3C-qPCR analysis of *IL20RA* interactions with RP11-10J5.1



3C-qPCR was carried out using the IL20RA_3 anchor primer with the lncRNA_1 test primer and NCR primers. For each set of primers a standard curve was generated using BAC control libraries spanning the region of interest. SYBR green qPCR was carried out in triplicate using 50ng of 3C library per reaction or a no-template (water) control. Triplicate 3C libraries were analysed for GM12878 (A) and Jurkats (B) and T-tests performed to determine if the relative interaction frequency was significantly different to the NCR. Only one library was analysed for the synovial fibroblasts (C) (error bars from technical replicates). Assays were also performed using 3C libraries from HapMap B-cell lines specific for the appropriate rs6927172 SNP genotype (D) - GM11994, GM10831, GM06993 (CG), GM12892, GM12707, GM10838 (CC), GM07037, GM10850, GM10858 (GG).

Summary of Results Section 2

- Statistically significant interactions identified through Capture Hi-C could be validated using 3C-qPCR in B-cells, T-cells and primary human synovial fibroblasts.
- RA associated SNPs in the 6q23 region interacted with *IL20RA, IFNGR1* and lncRNAs downstream of *TNFAIP3.*
- Interactions at the RA SNP region localised to a *Hin*dIII fragment containing two SNPs in LD with rs6920220 (rs6927172 and rs35926684) and another RA associated SNP (rs10499194).
- Interactions between the RA SNPs fragment and *IL20RA* and *IFNGR1* were shown to be correlated with carriage of the risk allele of rs6927172.
- Interactions between the RA SNPs fragment and lncRNAs were not genotype specific.
- Interactions between *IL20RA* and *TNFAIP3* and lncRNAs were not genotype specific.
- Interactions between *TNFAIP3* and lncRNAs were not genotype specific.
- Interactions not detected in Jurkats in the Capture Hi-C data analysis could be detected by 3C-qPCR.

Table 20: Summary of validated interactions

Anchor	Target	Validated in B-cells	Validated in T-cells	Validated in synovial fibroblasts	Genotype specific
IL20RA	rs6927172	✓	✓	✓	↑ Risk G allele
IFNGR1	rs6927172	✓	✓	✓	↑ Risk G allele
RP11-10J5.1	rs6927172	✓	✓	✓	X
RP11-240M16.1	rs6927172	✓	✓	✓	X
Downstream	rs6927172	✓	✓	✓	X
TNFAIP3	*psPTPN11*	✓	✓	✓	X
TNFAIP3	RP11-10J5.1	✓	✓	✓	X
TNFAIP3	RP11-240M16.1	✓	✓	✓	X
TNFAIP3	Y_RNA	✓	✓	X	X
TNFAIP3	*IL20RA*	✓	✓	✓	X
IL20RA	RP11-10J5.1	✓	✓	✓	X

5. Results Section 3

Analysis of regulatory protein binding by chromatin immunoprecipitation

5.1. Analysis of transcription factor binding by ChIP

Bioinformatic analysis of the 6q23 region containing RA associated SNPs showed evidence of transcription factor binding at that site, including NF-κB and BCL3. *TNFAIP3* is a regulator of NF-κB activity and BCL3 is a transcriptional co-activator that inhibits the nuclear translocation of the NF-κB p50 subunit in the cytoplasm and contributes to the regulation of transcription of NF-κB target genes in the nucleus.

ChIP assays for NF-κB subunits p50 and p65, and BCL3 were performed in GM12878 B-cells and Jurkat T-cells. ChIP was also carried out on B-LCLs carrying the three genotypes of the rs6927172 SNP to determine if there were any genotype-specific differences in transcription factor binding (non-risk CC = GM12892, GM07056, GM10843, GM10848, GM11993; heterozygous CG = GM12878, GM12875, GM12865; risk GG = GM10850, GM10858, GM12560). Analysis of ChIP enrichment was carried out by SYBR green qPCR, normalising to non-immunoprecipitated input chromatin.

5.1.1. Transcription factor ChIP assays

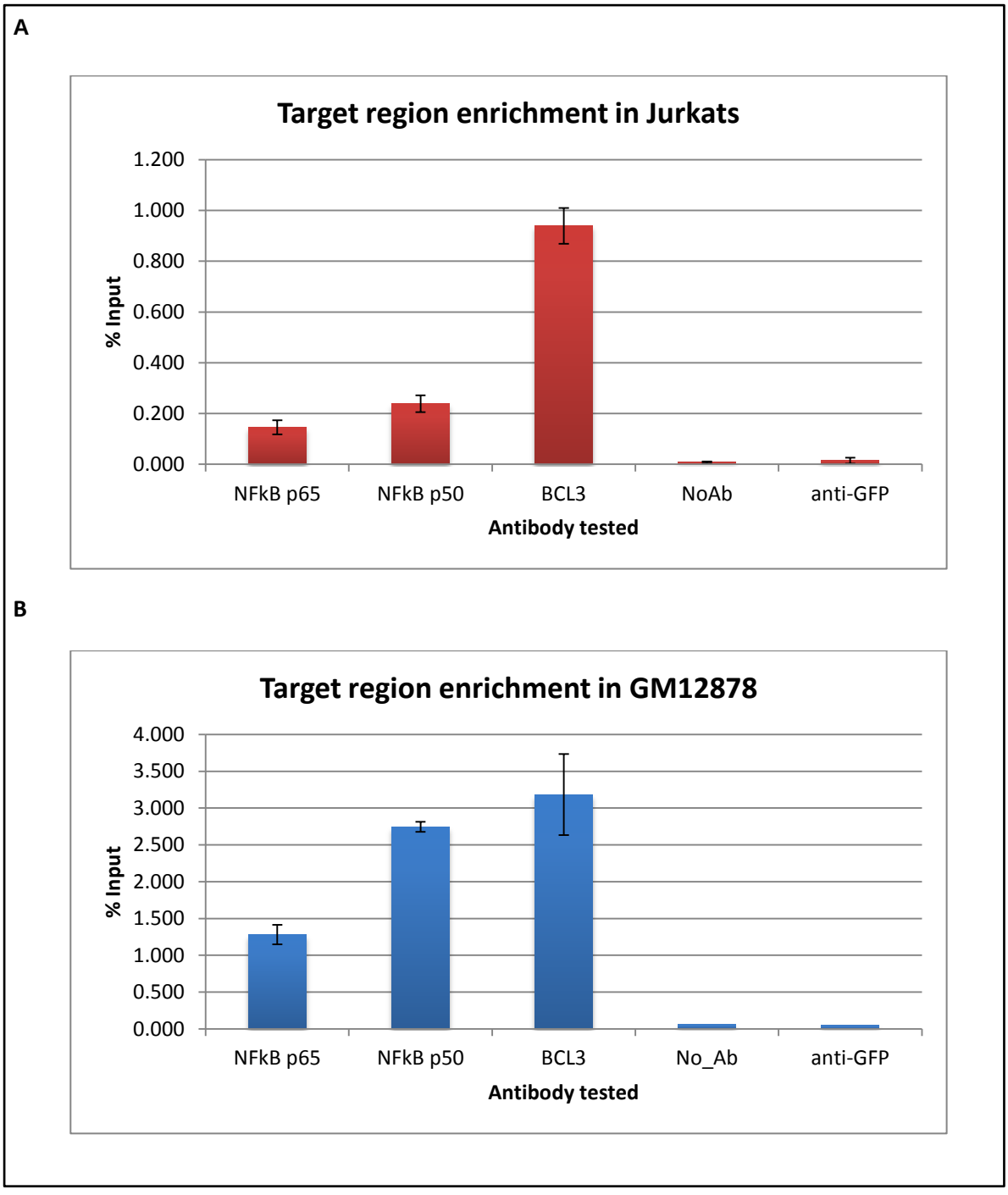
For both NF-κB and the BCL3 antibodies I had previously optimised the concentration of antibody used in the experiment, the type of magnetic beads, incubation duration and primers used for qPCR analysis.

Jurkat T-cells and LCLs containing the different rs6927172 genotypes were assayed for target region enrichment at the NF-κB and BCL3 transcription factor binding site using NF-κB p50, p65 and BCL3 antibodies. SYBR green qPCR was performed to detect enrichment of transcription factor binding at the target region compared to the input (non-IP'd) chromatin. The qPCR data was analysed as shown in section 2.7.3 to obtain the % Target enrichment values.

ChIP followed by SYBR green qPCR with Jurkat T-cells showed evidence of NF-κB p50, p65 and BCL3 binding at the target region containing the rs6927172 SNP (Figure 57A). The GM12878 LCLs also showed evidence of NF-κB p50, p65 and BCL3 transcription factor binding at the target region containing the rs6927172 SNP (Figure 57B).

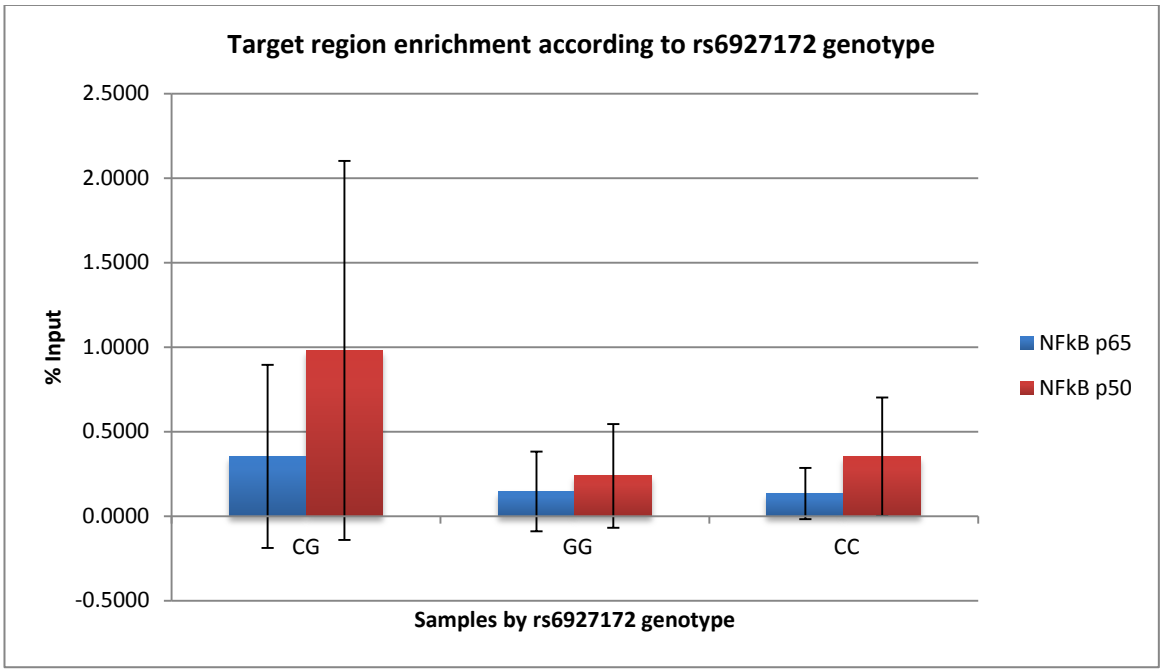
LCLs containing the three different genotypes for rs6927172 did not show any significant differences in NF-κB p50 or p65 binding at the target region and showed a large degree of variability between different cells containing the same genotype (Figure 58 and Appendix Figures 68-70). Genotype-specific ChIP was not carried out for the BCL3 antibody. Genotype-specific assays were unable to be carried out on the Jurkat T-cells as it is just the one cell-line, which is heterozygous for the rs6927172 SNP. However, an allele-specific TaqMan assay, carried out by others in my group, using probes targeting the rs6927172 SNP in Jurkats has shown an increase in NF-κB p65 enrichment in the presence of the risk allele (see Appendix Figure 71 - unpublished data, manuscript in preparation).

Figure 57: Transcription factor binding in B-cell and T-cell lines at the 6q23 rs6927172 RA SNP target region



Jurkat (A) and GM12878 (B) cell lines were analysed for transcription factor binding at the rs6927172 target region using ChIP grade antibodies. Each ChIP was carried out on three technical replicates for each cell line. The data represents the average % Input results from the three ChIP replicates; error bars are +/- St.Dev. qPCR using SYBR green on a QuantStudio 12K Flex instrument was carried out in triplicate using primers specific for the target region, a positive control region and negative control region (see Appendix Figure 72 for positive and negative control qPCR chart).

Figure 58: Summary of target region enrichment in NF-κB ChIP assays according to rs6927172 Genotype



B-lymphoblastoid cell lines containing the different rs6927172 genotypes (non-risk CC = GM12892, GM07056, GM10843, GM10848, GM11993; heterozygous CG = GM12878, GM12875, GM12865; risk GG = GM10850, GM10858, GM12560) were tested for NF-kB p50 and p65 transcription factor binding at the rs6927172 target region. Each ChIP was carried out in triplicate along with a no antibody control. SYBR green qPCR was carried out in triplicate on a QuantStudio 12K Flex instrument using primers specific for the target region, a positive control region and negative control region (see Appendix Figure 72 for positive and negative control qPCR chart). ChIP samples were normalised to the non-IP'd Input sample and the no-antibody control was used to determine the level of non-specific, background binding which was subtracted off the sample values. The data shown represents the average % Input results from the samples tested for each genotype; error bars are +/- St.Dev.

5.1.2. Statistical analysis of genotype-specific ChIP enrichment

Firstly, a summary table was produced for each antibody (Tables 21-23). T-tests were carried out to determine if there was a statistically significant difference in NF-κB binding between genotypes ($p<0.05$). The data used is included in the Appendix Tables 53-56.

No significant differences were detected between rs6927172 genotypes in B-LCLs for either NF-kB p65 or NF-kB p50.

Table 21: Summary table of NF-κB p65 results

rs6927172 genotype "CC" (Major allele)					
Variable	Samples	Average	S.E.M.	Min	Max
ChIP input %	15	0.1451	0.24	-0.01	0.46
rs6927172 genotype "CG"					
ChIP input %	9	0.3531	0.54	0.00	0.98
rs6927172 genotype "GG" (Minor allele)					
ChIP input %	9	0.1342	0.15	0.03	0.31
T-Test					
CG vs CC	0.44				
CC vs GG	0.33				

Table 22: Summary table of NF-κB p50 results

rs6927172 genotype "CC" (Major allele)					
Variable	Samples	Average	S.E.M.	Min	Max
ChIP input %	15	0.2381	0.31	-0.08	0.71
rs6927172 genotype "CG"					
ChIP input %	9	0.9798	1.12	0.31	2.27
rs6927172 genotype "GG" (Minor allele)					
ChIP input %	9	0.3509	0.35	0.05	0.74
T-Test					
CG vs CC	0.30				
CC vs GG	0.17				

Table 23: Summary table of BCL3 results

rs6927172 genotype "CG"					
Variable	Samples	Average	S.E.M.	Min	Max
ChIP input %	3	3.174	0.067	2.79	3.57

Conclusions:

Statistical analysis showed that there was no significant difference in target region enrichment between the different rs6927172 alleles for NF-κB p50 or NF-κB p65 transcription factors in the B-cell lines analysed.

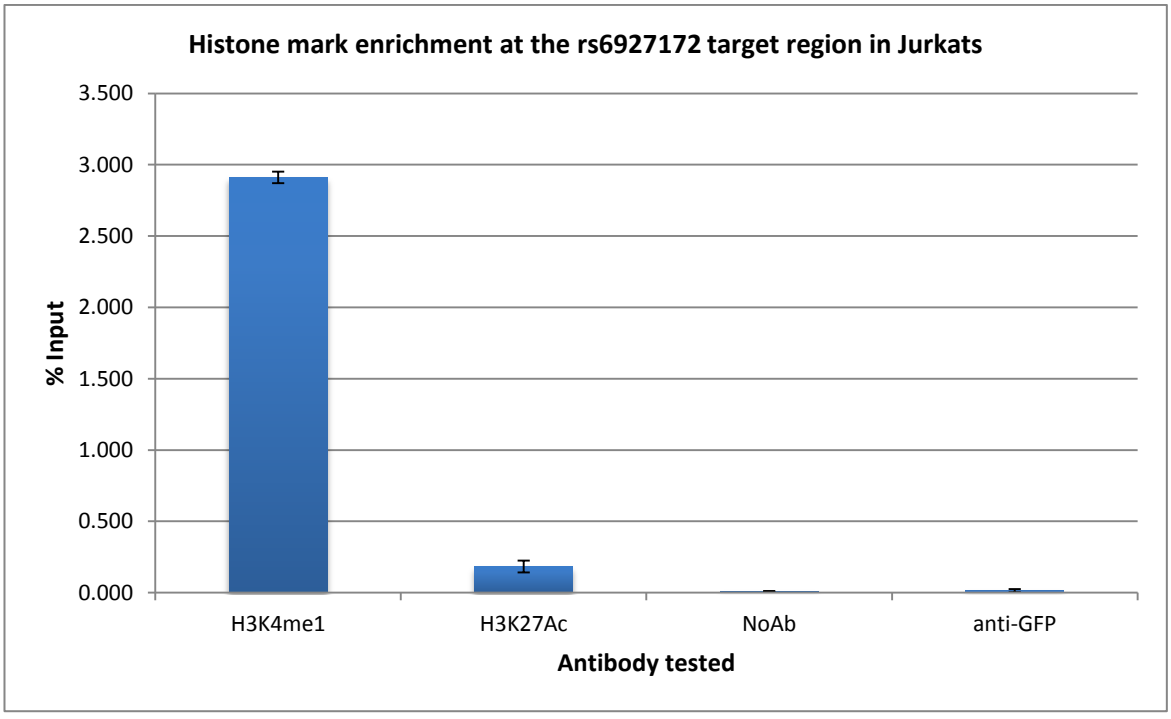
5.2. ChIP assays for enrichment of histone marks

Bioinformatics analysis provided evidence that the RA associated SNPs lie in a region of open chromatin containing histone marks indicative of enhancer activity. Enrichment of the active enhancer histone marks H3K4me1 and H3K27ac was analysed by ChIP in B-LCLs and Jurkats.

ChIP with antibodies for histone marks suggested, in both B-cells and T-cells, that the RA associated SNPs lied in an active enhancer region (Figure 59 and Figure 60). B-cells containing the non-risk allele showed evidence of increased binding of enhancer marks which was statistically significant ($p<0.05$ for CC vs CG and CC vs GG) (Figure 60).

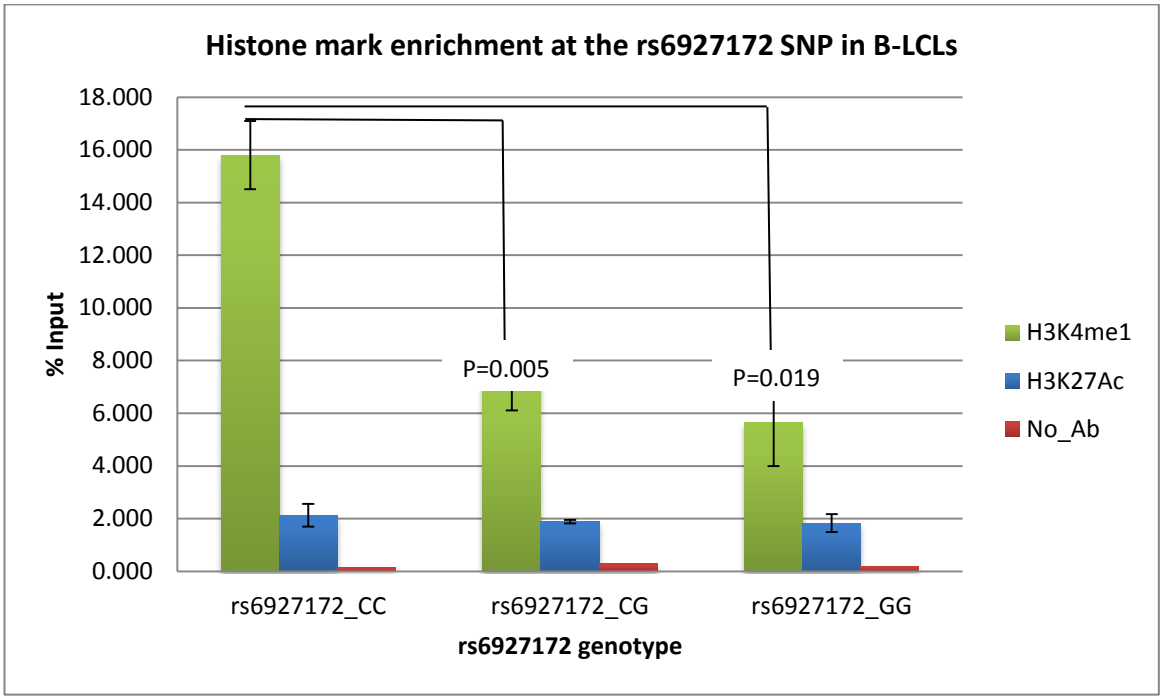
Allele-specific TaqMan assays, carried out by others in my group, using probes targeting the rs6927172 SNP in Jurkats have shown a modest increase in histone mark enrichment in the presence of the risk allele (see Appendix Figure 71 - unpublished data, manuscript in preparation), which is the opposite effect to what is shown in the B-cells here.

Figure 59: Enrichment of histone marks at the rs6927172 target region in Jurkats



Enrichment of the histone marks H3K4me1 and H3K27ac at the rs6927172 target region was assessed in Jurkats. Each ChIP was carried out in triplicate along with a no antibody control. SYBR green qPCR was carried out in triplicate on a QuantStudio 12K Flex instrument using primers specific for the target region, a positive control region and negative control region (see Appendix Figure 72 for positive and negative control qPCR chart). ChIP samples were normalised to the non-IP'd Input sample. The data shown represents the average % Input results from the ChIP replicates; error bars are +/- St.Dev.

Figure 60: Enrichment of histone marks at the rs6927172 target region in B-LCLs



B-lymphoblastoid cell lines containing the different rs6927172 genotypes (CC – GM06985, GM10838, GM12892; CG – GM10831, GM11994, GM06993; GG – GM10858, GM10850, GM07037) were tested for H3K4me1 and H3K27ac histone mark enrichment at the rs6927172 target region. Each ChIP was carried out in triplicate along with a no antibody control. SYBR green qPCR was carried out in triplicate on a QuantStudio 12K Flex instrument using primers specific for the target region, a positive control region and negative control region (see Appendix Figure 72 for positive and negative control qPCR chart). ChIP samples were normalised to the non-IP'd Input sample. The data shown represents the average % Input results from the samples tested for each genotype; error bars are +/- St.Dev.

Summary of Results Section 3

- ChIP with antibodies for histone marks H3K4me1 and H3K27ac suggested, in both B-cells and T-cells, that the RA associated SNPs lied in an active enhancer region.

- B-cells containing the non-risk allele of rs6927172 showed evidence of increased binding of enhancer marks H3K4me1 and H3K27ac which was statistically significant.

- Jurkat T-cells showed evidence of NF-κB p50, p65 and BCL3 binding at the target region containing the rs6927172 SNP.

- GM12878 LCLs (heterozygous for rs6927172) showed evidence of NF-κB p50, p65 and BCL3 transcription factor binding at the target region containing the rs6927172 SNP.

- No genotype-specific differences in NF-κB binding at the SNP target region were detectable in B-LCLs.

- Additional experiments by others in my group, using TaqMan probes targeting the rs6927172 SNP in Jurkats, have shown an increase in NF-κB p65, H3K4me1 and H3K27ac enrichment in the presence of the risk allele.

6. Discussion

6.1. Summary of findings

GWAS have identified many SNP variants associated with the onset of RA. The challenge post-GWAS for all diseases, and the ultimate aim of this project was to identify the causal variants in an RA associated locus, to find which genes are being affected by the variants and to elucidate the mechanism by which these variants modify gene function.

Capture Hi-C was used in this study to investigate long-range interactions between disease associated regions and gene promoters using two complementary solution capture hybridisations. Quality control of Capture Hi-C libraries, both pre-capture and post-capture, showed that the libraries were of a consistently high standard giving confidence that the experiments were generating reproducible, high quality data for analysis. Analysis of Capture Hi-C data revealed numerous long-range interactions implicating novel candidate genes that had not previously been considered in GWAS and showed that interactions involving multiple genetic loci could have common interaction targets.

Analysis of the 6q23 region revealed that SNPs associated with RA interacted with novel candidate genes, *IL20RA, IFNGR1* and also with regulatory lncRNAs downstream of *TNFAIP3.* Within the 6q23 region, statistically significant interactions identified through Capture Hi-C were validated using 3C-qPCR in B-cells, T-cells and primary human synovial fibroblasts and provided evidence that interactions involving the RA SNP region localised to a *Hin*dIII fragment containing two SNPs in LD with rs6920220 (rs6927172 and rs35926684) and another RA associated SNP (rs10499194). Bioinformatic analysis provided evidence that the rs6927172 SNP was the most likely regulatory SNP in the 6q23 intergenic region, and interestingly, interactions between the RA SNPs fragment and *IL20RA* and *IFNGR1* were shown to be correlated with carriage of the risk allele of rs6927172.

To further investigate the potential functional role for the rs6927172 SNP, ChIP was used to investigate DNA-protein interactions within the SNP target region. Bioinformatics evidence suggested the SNP lied in a region of enhancer activity, therefore ChIP with antibodies for histone marks H3K4me1 and H3K27ac was performed which indicated, in both B-cells and T-cells, that the RA associated SNPs did indeed lie in an active enhancer region. B-cells containing the non-risk allele of rs6927172 showed evidence of increased binding of enhancer marks, however in T-cells presence of the risk allele suggested decreased binding of enhancer marks.

Analysis of transcription factor binding at the SNP target region showed evidence of NF-κB p50, p65 and BCL3 binding in both B-cells and T-cells. No genotype specific differences were detected in B-cells but carriage of the risk allele in T-cells suggested decreased binding of transcription factors at the SNP target site, providing further evidence that the SNPs could potentially alter target gene expression, possibly through altered binding of regulatory proteins.

In summary, chromosome conformation capture and ChIP experiments have revealed that the spatial organisation of the chromatin at the 6q23 region is complex, bringing together several genes with key roles in the immune response, including *IL20RA*, *IFNGR1* and *TNFAIP3*, along with regulatory elements containing SNPs associated with different autoimmune diseases. Also, evidence obtained from Capture Hi-C, targeted 3C-qPCR and bioinformatics all suggests that the rs6927172 SNP, which is in perfect LD with the GWAS index SNP rs6920220, is the most likely functional SNP in the 6q23 region.

Taken together, the results presented in this thesis suggest that the mechanism by which the risk allele of rs6927172 alters expression of genes such as *IL20RA, IFNGR1* and *TNFAIP3* may be mediated by an increased regulatory activity and augmented transcription factor binding.

6.2. Background

This project originated from the RA GWAS and genetic fine-mapping studies, co-ordinated by the Manchester group (Eyre *et al.* 2012; WTCCC 2007). Genetic fine-mapping studies can only go so far in identifying the SNPs and genes likely to be causal in disease susceptibility. Fine-mapping often implicates a number of independently associated genetic signals within a locus, each in strong linkage disequilibrium (LD) (Reich *et al.* 2001) with a number of other – potentially equally likely causal – variants. LD is the occurrence of combinations of alleles or genetic markers in a population more or less often than would be expected from a random formation of haplotypes from alleles based on their frequencies. This means that, although there is strong association with one SNP, other SNPs highly correlated with this variant will be equally associated and may be more likely to have a functional effect.

In addition, most of the variants identified from GWAS are in intergenic, non-protein coding regions of the genome, therefore the causal variants within RA associated regions may be acting by influencing gene regulation, possibly through physical contact with distal target genes and/or alteration of binding of regulatory proteins. It is therefore not trivial to assign genes and causal SNPs to the genetic signals emerging from GWAS efforts.

The laboratory in Manchester I am sited in has traditionally focussed on the genetic association in rheumatic diseases. It is now transitioning into the translation of GWAS findings, using a multi-discipline, functional biology approach, therefore a large part of my PhD was the introduction of new molecular techniques into the laboratory in a robust and reproducible way.

6.3. The post-GWAS landscape

To translate the findings of GWAS the challenge now is to link the association signal to gene/s, cell type and stimulus in order to identify disease pathways and new therapeutic targets or to re-direct existing therapies.

More than 90% of SNPs associated with autoimmune disease, identified through GWAS, are located in intergenic, non-coding regions of the genome (Farh *et al.* 2015; Maurano *et al.* 2012; Maurano *et al.* 2015; Ricano-Ponce *et al.* 2016) meaning that their roles are poorly understood. Multiple lines of evidence suggest a regulatory role for intergenic SNPs. Almost 8% of variants map to promoter histone marks (Ricano-Ponce *et al.* 2013), approximately 60% map to enhancer marks (Farh *et al.* 2015), 32% are in DNaseI hypersensitivity sites, and 10-20% map to protein binding sites, potentially altering transcription factor binding ability and, therefore, gene expression (Farh *et al.* 2015; Ricano-Ponce *et al.* 2013). However, some of these estimates are based on incomplete datasets from experiments such as ChIP-seq, carried out on only a limited number of cell types and stimulatory conditions so these numbers may not be accurate and more work is required to systematically characterise functional annotation.

Recently, the FANTOM5 (Functional Annotation of the Mammalian Genome 5) project (Andersson *et al.* 2014; Forrest *et al.* 2014) have published comprehensive maps of transcription factors, enhancers, promoters and regulatory networks in 432 primary cell types, 135 tissues and 241 cell-lines. Enhancers defined in FANTOM5 were shown to be enriched in GWAS SNPs (Andersson *et al.* 2014). Autoimmune associated loci have been shown to be enriched in immune-cell enhancers in a number of studies (Farh *et al.* 2015; Hawkins *et al.* 2013; Maurano *et al.* 2012). Disease-associated SNPs are enriched in CD4+ T-cells, CD8+ T-cells, and B-cell enhancers (Farh *et al.* 2015) and in RA, CD4+ T_{REG} -cells show enrichment of a histone mark that is characteristic of active gene regulation, H3K4me3 (Trynka *et al.* 2013).

GWAS SNPs have also been shown to be enriched in super-enhancers (Hnisz *et al.* 2013; Parker *et al.* 2013), which are comprised of a complex of enhancers, occupied by key transcription factors and co-activators. Super-enhancers are associated with driving the expression of genes controlling cell-type identity (Hnisz *et al.* 2013; Pott *et al.* 2015). Indeed, GWAS SNPs have been found to be enriched 7.5-fold in CD4+ T-cell super-enhancers (Farh *et al.* 2015), providing further evidence for a regulatory role of intergenic SNPs and also for cell-type specific gene expression patterns.

Studying the correct cell-type is imperative to move forward in the post-GWAS era in order to identify the most biologically relevant genes contributing to a particular disease. For example, Maurano *et al* (Maurano *et al.* 2012) showed that DNaseI-HS sites that defined Th17 and Th1 cells are enriched in variants associated with Crohn's disease, and that this marker of open, transcriptionally active chromatin was enriched for associated variants of multiple sclerosis in both CD4+ T-cells from umbilical cord blood and CD19+/CD20+ B-cells. As previously mentioned, this type of evidence, overlaying markers of cell type specific enhancers such as H3K3me3, with

GWAS identified variants has shown that CD4+ T-cells are of particular importance in RA (Diogo *et al.* 2014; Trynka *et al.* 2013). These studies demonstrated that RA associated variants were statistically significantly enriched in regions of the genome that are preferentially open and active in CD4+ T-cells, based on epigenetic marks.

Most of the post-GWAS efforts in order to link SNPs to genes have been centred around eQTL analysis. It has been shown that intergenic SNPs can affect the expression of nearby (<1Mb away) transcripts (*cis*-eQTLs) (Kumar *et al.* 2014; Ricano-Ponce *et al.* 2013; Wright *et al.* 2014). Initial eQTL studies were carried out in B-LCLs or PBMCs (Stranger *et al.* 2012; Wright *et al.* 2014) but now it has been recognised that it is important to perform these studies in specific cell types as it has been shown that many eQTLs are cell-type, tissue and stimulus specific (Dimas *et al.* 2009; Edwards *et al.* 2013; Maurano *et al.* 2012).

As part of the ImmVar project, CD4+ T-cells were analysed under unstimulated and active conditions (differentiated into T_H17 cells) to profile gene expression under different conditions (Ye *et al.* 2014). This study found that *cis* genetic affects accounted for around 25% of the heritability of gene expression, almost half of the genes studied had a cis-eQTL within 1Mb, but a third of these were stimulation or time specific. Context-specific eQTLs which were enriched for disease associated loci have been also been demonstrated in primary human monocytes which had been stimulated using two different stimuli for different durations (Fairfax *et al.* 2014). Here again it was demonstrated that most genes show evidence of an eQTL (83.7%), but that these were split equally between activation and stimulation states, for example 25% were found in naïve cells, 21% found when stimulating cells for two hours with LPS and 28% found after IFN-γ stimulation. This illustrates how the effect of associated variants can change based on context, such that it is imperative to determine the effect of these variants in cell types and stimulatory conditions that are relevant to the disease. In RA, genetic evidence based on the epigenetic marks overlaying associated SNPs and the enrichment around immune genes, strongly implicates CD4+ T-cells in disease and therefore this is why we chose to carry out experiments in T-cells.

Most eQTL studies have been carried out on cells from healthy volunteers so it would be very interesting to see eQTL data from patients in order to explore the inflammatory environment. Large international consortiums such as GTEx (GTEx Project. 2013) are working towards a more complete picture of eQTLs in many more cell types and conditions, which will be an invaluable resource.

Multiple SNPs are often found clustered in disease associated regions and are in close LD, therefore it is a considerable challenge to identify the precise causal SNPs in these regions (Farh *et al.* 2015). In order to prioritise potential causal SNPs within an LD block, computational approaches have been developed. PICS (probabilistic identification of causal SNPs) is an algorithm which estimates the probability that an individual SNP is a causal variant given the haplotype structure and observed pattern of association at the locus (Farh *et al.* 2015). Using the

PICS algorithm found that index SNPs reported in GWAS had only 5% chance of being the causal SNP.

Another approach to prioritise causal SNPs is the use of credible sets as used in our Capture Hi-C study (Wallace *et al.* 2015). Prediction of the causal SNP after fine-mapping can be carried out using a set of SNPs accounting for 95% or 99% of the probability. If the true causal SNP has been fine mapped it will be contained within the relevant credible SNP set. For example, credible sets analysis carried out by others in our group for the 6q23 region chr6:137895629-138125334 identified 802 SNPs, but this region was narrowed down to chr6:137973832-138006504, with the most likely causal SNPs in the region being rs17264332, rs11757201, rs6920220, rs6927172 and rs928722.

It is clear that the relationship between SNPs and genes is incredibly complex and statistical approaches cannot provide all the answers. Computational approaches can only provide predictions that need to be experimentally tested, therefore to continue research in the post-GWAS era, functional studies within the laboratory are of the utmost importance in order to uncover the molecular mechanisms involved in the disease process.

An example of how a GWAS SNP can translate to phenotype was demonstrated in the pioneering study by Musunuru *et al* (Musunuru *et al.* 2010) which investigated a locus on chromosome 1p13 associated with cholesterol and myocardial infarction in humans. The SNPs were found to be eQTLs, but only in liver tissue. Luciferase reporter assays and site-directed mutagenesis provided compelling evidence as to the causal variant, whereby the identified SNP created a transcription factor binding site which was shown to alter the expression of *SORT1*, and this was linked to plasma levels of cholesterol, potentially altering the risk for myocardial infarction. This was one of the first studies that went from associated region, through SNP, gene, mechanism and function to highlight a candidate gene and tissue from GWAS data.

Further examples of how GWAS SNPs have been linked to specific genes include a variant associated with renal cell carcinoma which resulted in impaired binding and function of hypoxia inducible factor at a novel enhancer of *CCND1* (Schodel *et al.* 2012). An intergenic variant associated with foetal haemoglobin levels could disrupt transcription factor binding in an erythroid-specific enhancer resulting in the reduced expression of *BCL11A* (Bauer *et al*. 2013). Variants associated with colorectal and prostate cancer have been shown to modulate transcription factor binding at enhancer elements through long-range looping interactions with *Myc* and *SOX9* (Pomerantz *et al.* 2009; Schodel *et al.* 2012; Zhang *et al.* 2012c).

Breast cancer research has also shown some compelling evidence linking intergenic SNPs to phenotype (French *et al.* 2013; Meyer *et al.* 2013). Functional variants at the 11q13 risk locus were investigated using a variety of techniques including eQTL analysis, ChIA-PET, 3C, luciferase reporter assays, siRNA knockdown, EMSA and ChIP (French *et al.* 2013). Using these functional assays, three SNPs were found to be very strong candidates for having a causal effect by directly

affecting a gene, cyclin D1 which is considered an oncogene. A similar approach was used to investigate variants within the 10q26 breast cancer locus (Meyer *et al.* 2013). Genetic fine-mapping, DNase hypersensitivity data and EMSA identified three putative causal SNPs. ChIP showed preferential transcription factor binding to the risk variant and 3C assays demonstrated that the risk region could interact with a gene promoter which was the most likely target gene of the risk region.

There are some recent functional studies in autoimmunity loci which use some of the techniques used in my study. An example is the investigation of the 16p13 region associated with T1D and MS (Davison *et al.* 2012). This study used eQTL analysis, which found that *DEXI* expression was correlated with many SNPs in *CLEC16A,* and 3C analysis, which identified a 150kb looping interaction between *CLEC16A* and the promoter of *DEXI*, suggesting that variants within *CLEC16A* could regulate the expression of *DEXI*. This study showed the ability of functional studies to identify potential novel candidate genes within a locus.

Another example of a functional study focused on the TT>A polymorphism in the 6q23 locus downstream of *TNFAIP3*, which has been associated with SLE (Wang *et al.* 2013), and is correlated with reduced expression of A20 in patients with the risk allele. Functional analysis of this TT>A variant was carried out using EMSA, reporter assays, ChIP and 3C in HapMap LCLs and showed that the variant could physically interact with the promoter of *TNFAIP3* through long-range looping, and A20 expression could be affected through inefficient delivery of NF-κB to the *TNFAIP3* promoter. This study demonstrated a variety of ways in which the function of a SNP can be investigated.

My project looked to use the techniques employed in these studies, and to build upon them with several improvements. For example, I was able to employ the latest expanded bioinformatics data, from sources such as ENCODE, Epigenetics Roadmap and Blueprint, that have made enormous strides in terms of depth of data and the number of cell types analysed. The wealth of data, including methylation, expression, histone marks, chromatin accessibility and interaction, on hundreds of primary, stimulated and cell lines, was simply not available to inform these earlier studies. I was able to incorporate more expression data, for example from the ImmVar project, but also from other studies in primary cells, stimulated cells and more individual cell types. All this data increases our understanding of the mechanisms of gene regulation, and helps define better hypotheses to test with laboratory experiments. Also my primary research technique (Capture Hi-C) is a major advancement on the interaction data available to previous studies. The fact that my approach was hypothesis free, determining all interactions with a selected associated enhancer, and generated high resolution data for the interactions meant my work really adds to the knowledge in this field and develops the area of GWAS-enhancer interactions.

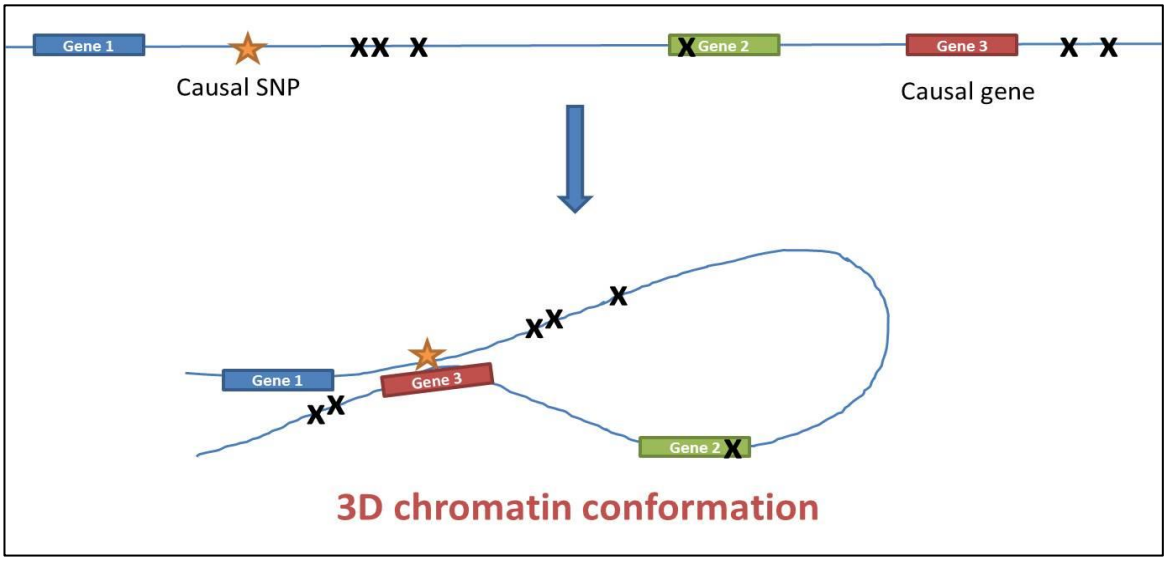
6.4. Post-GWAS investigation of RA loci

6.4.1. Investigation of chromatin folding

Potential causal variants often lie some distance from genes when looking at the linear genome, therefore the 3-D folding of chromatin could dictate how SNPs affect regulatory activity, illustrated in Figure 61. There is well established evidence that chromatin folding can bring genomic regions that are far apart into close proximity (Davison *et al.* 2012; Pomerantz *et al.* 2009; Zhang *et al.* 2012d) to play a role in transcriptional regulation (Fraser *et al.* 2007; Smallwood *et al.* 2013).

Chromosome conformation capture (3C) technologies have developed over the last few years from a simple technique to identify one-to-one interactions, to a powerful tool enabling a genome-wide view of the chromatin interaction landscape. This makes these techniques an ideal way of following up on GWAS associations.

Figure 61: 3-D chromatin conformation can bring distant elements together



Through looping interactions, a single promoter can interact with multiple enhancers leading to complex patterns of gene expression which are cell-type specific (Bulger *et al.* 2011; Dixon *et al.* 2012). Promoter-enhancer interactions occur preferentially within topologically associated domains (TADs) (Jin *et al.* 2013) creating large domains which contribute to specific gene expression profiles (Mifsud *et al.* 2015; Symmons *et al.* 2014).

As previously mentioned, 3C has successfully been used to interrogate long-range interactions in regions associated with T1D, MS and SLE making 3C a valid technique to use in this study.

Why Capture Hi-C?

Identifying long-range interactions between disease-associated SNPs and distal genes can give confidence that the correct gene has been identified. Previous studies have focused on either one-to-one interactions (3C), interactions between one target and many potential interacting partners (4C) or interactions involving many targets and many potential partners (5C) (Sanyal *et al.* 2012; Sexton *et al.* 2012a; Simonis *et al.* 2006; Splinter *et al.* 2012). However, none of these techniques are genome wide and require prior knowledge of the target regions in order to design primers.

Hi-C has been mainly used to study the 3D organisation of the mouse and human genomes (Dixon *et al.* 2012; Lieberman-Aiden *et al.* 2009), and has been used to study the dynamics of promoter-enhancer signalling (Jin *et al.* 2013). However, the data obtained using Hi-C makes it difficult to identify specific interactions because to confidently call interactions they must be seen as higher than background. Since Hi-C interrogates all possible interactions it requires an enormous amount of sequencing depth to generate the number of interactions required to confidently call any individual interaction. More usually, interactions have to be collapsed into 'bins' to gain sufficient sequencing depth which reduces the effective resolution of the experiment and only allows interactions between large 'bins' to be called.

The recent development of Capture Hi-C (CHi-C) allows a genome-wide view of interactions by combining Hi-C with a solution capture hybridisation step followed by NGS (Dryden *et al.* 2014; Jager *et al.* 2015; Mifsud *et al.* 2015; Schoenfelder *et al.* 2015). This method massively increases the number of 'on target' reads, that is, the regions the investigators are interested in. By using CHi-C it is possible to resolve interactions to a single restriction fragment level, around 4kb. This resolution is sufficient to link enhancers to gene promoters, something difficult, or almost impossible with Hi-C alone. Therefore CHi-C combines the 'hypothesis free' advantage of Hi-C with the resolution of 3C/4C, making it currently the stand out method for interrogating GWAS enhancer interactions. CHi-C was therefore used in this study with the aim of pinpointing genes involved in disease which could then potentially suggest ætiopathological pathways in relevant immune cell types.

The important limitations of using Hi-C were neatly demonstrated in work I have carried out in the lab involved sequencing Hi-C libraries from both LCLs and Jurkat cell-lines. The data obtained was indeed low resolution, requiring extensive binning of restriction fragments. The baited approach I finally used overcame this limitation in resolution by targeting specific regions, whilst still allowing a genome-wide interaction profile through the use of Hi-C libraries. Capture Hi-C protocols became available during the early stages of my PhD therefore, a key focus during the first year of this project was to generate high-quality Hi-C libraries in both GM12878 B-LCL and Jurkat T-cell lines for use in Capture Hi-C experiments.

6.4.2. Library generation for Capture Hi-C

Generating Capture Hi-C libraries is complex and time-consuming

The number of potential DNA interactions happening in a cell is extremely high, up to 10^{11} possible unique interactions, however only two interactions per fragment can be detected from a single cell (Belton *et al.* 2012). It is therefore necessary to start the experiment with a large cell number, usually 2-3x10^7 or more cells, in order to make a complex library capturing as many interactions as possible. Therefore, to ensure good library complexity, biological replicate Hi-C libraries from LCLs and Jurkats were produced from 5-6x10^7 cells.

This is of particular consideration when precious samples or rare cell populations are being investigated as it may be unlikely that a high number of cells could be obtained. Hi-C libraries can be made from 1-5x10^6 cells but the complexity may be low (Belton *et al.* 2012) and such a low cell number may not yield enough library to perform a capture experiment which requires a large amount of input. Indeed, a cell-number titration I carried out generated successful libraries with as low as 1x10^7 cells but whilst there would be enough material to directly sequence, there was not enough material to use in a capture when fewer than 2x10^7 cells were used (in-house data generated by myself). In the future, the use of lower cell numbers to generate a library for capture experiments will hopefully be possible as improved protocols evolve. For example, the vast majority of the DNA material is lost through extensive clean-up steps and enzymatic reactions used in the library generation as discussed below. Technologies to simplify library preparation, or techniques that require less input for sample capture, would help reduce the number of input cells required. This is particular important as we move towards trying to determine the functional interactions in more homogeneous populations of primary cells, for example T_{REG} cells, where obtaining the numbers of cells for starting material currently required is not feasible.

Critical steps and quality control

The Hi-C protocol consists of many stages that are of critical importance to ensure the library will yield good quality sequencing results. The crosslinking of interactions is the first critical stage. If the cells are under crosslinked interactions will be missed, whilst over-crosslinking can make the chromatin inaccessible to restriction enzymes resulting in reduced digestion efficiency, increasing the amount of random collisions leading to higher background (Duan *et al.* 2012). Here, cells were prepared for crosslinking by using serum-free culture medium because serum has the potential to affect crosslinking efficiency by sequestering the formaldehyde in the culture medium, bound to serum proteins, rather than inside the cells (Belton *et al.* 2012).

The restriction enzyme choice is an important part of the Hi-C experimental design as the results obtained from sequencing are based on identifying ligation products formed by interacting fragments joined at restriction sites. Restriction enzymes which recognise 6-bp DNA sequences, such as *Hin*dIII, *Eco*RI, *Bgl*II, and *Nco*I are popular choices because they have been shown to produce good-quality data (Duan *et al.* 2012). The average DNA length, and therefore, resolution

obtained by using 6-cutters is around 4kb. To increase the resolution of the experiment, restriction enzymes which recognise a 4bp DNA sequence (Simonis *et al.* 2007), such as *Dpn*I (Comet *et al.* 2011) or *Mbo*I (Rao *et al.* 2014) which cut every 256bp, can be used. *Hin*dIII was chosen to generate the Hi-C libraries, being routinely used in available protocols with published PCR primers readily available for sample QC, giving confidence in the library quality. *Hin*dIII recognises the sequence 5'-AAGCTT-3' and leaves an overhang of 5'-AGCT-3' after DNA cleavage where biotin can be incorporated when the overhang is filled in using dNTPs. In the original Hi-C protocol (Lieberman-Aiden *et al.* 2009) biotin-14-dCTP was used, but the protocol used for generating my Hi-C libraries used biotin-14-d**A**TP which has been shown to improve the biotinylation reaction efficiency (Dryden *et al.* 2014).

The principle of proximity ligation in 3C protocols assumes that ligation reactions using a dilute solution favours the ligation of cross-linked DNA fragments in preference to ligation of individual fragments (Gavrilov *et al.* 2013; Lieberman-Aiden *et al.* 2009). However, other groups have shown that dilution is not necessary to preserve 3C profiles (Comet *et al.* 2011) and that most of the information comes from ligations occurring in the insoluble fraction of chromatin within the nucleus (Gavrilov *et al.* 2013). Traditionally, the protocol I used specified dilute conditions for the ligation reactions in order to minimise background interactions.

Prior to the biotinylation step the original Hi-C protocol lysed the nuclei so that the ligation step was carried out in solution, using a high volume to increase the likelihood of the formation of true interactions occurring as opposed to random collisions (Lieberman-Aiden *et al.* 2009). The libraries I produced used an improved protocol developed at the Babraham Institute which carried out the ligation step in intact nuclei, which has been shown conclusively in a direct comparison between in-solution and in-nucleus protocols to considerably reduce background noise and improve reproducibility between experiments (Nagano *et al.* 2013; Nagano *et al.* 2015).

The Hi-C protocol is costly in terms of time and reagents, so it is important to carry out quality control to detect any problems with the library at an early stage. For example, a large amount of low molecular weight products when the libraries are run on an agarose gel can indicate library degradation which can be caused by heating of the sample during lysis or the action of proteases and endogenous nucleases (Belton *et al.* 2012). To prevent heating of the samples and protease activity, cell lysis was carried out on ice using buffer containing protease inhibitors. The resulting libraries consistently showed no evidence of degradation.

As a control, 3C libraries were generated at the same time as the Hi-C libraries by omitting the biotinylation step of the protocol (Belton *et al.* 2012). These libraries were used as a comparison in the PCR digest assay to assess biotinylation and ligation reaction efficiency. Published PCR primers were used to amplify ligation products formed from two adjacent restriction fragments, which were subsequently digested using *Hin*dIII and *Nhe*I. The biotinylation efficiency in Hi-C experiments can be between 5-30% as it is a blunt-ended ligation, so it is not very efficient (Belton

et al. 2012). The PCR digest assay can show if the biotinylation and fill-in and ligation stages were successful because the Hi-C libraries will only digest with *Nhe*I if this stage has worked.

The QC steps carried out on the LCL and Jurkat libraries all indicated that the libraries were of high quality, demonstrated known control short-range and long-range interactions, and were proven to be Hi-C libraries by successful PCR digest assay. Therefore, the material could be used to complete the Hi-C libraries ready for capture and sequencing. The remaining steps of the library generation involved removal of biotin from non-ligated ends, chromatin shearing, size selection, ligation of sequencing adapters, biotin-streptavidin pulldown of Hi-C ligation products and PCR enrichment.

The final steps of the library preparation are where a majority of the starting material is lost. Biotin removal from non-ligated ends is carried out on 40µg of Hi-C library, however, the expected yield following size selection is approximately 10µg so 75% of the library is lost. For all the libraries generated, the full 40µg was carried through to the biotin removal step, as a high yield of library was obtained, resulting in the expected yield of approximately 10µg after these steps.

The material loss arises from a number of experimental steps. Firstly, the biotinylation reaction and ligation reactions are not 100% efficient meaning that not every ligation junction gets biotinylated, so at the pulldown stage not all potential ligation products can be captured. Secondly, there are a lot of wash steps in the protocol because high levels of contaminants such as salts and enzymes can be left in the samples following ligation under dilute conditions and some material will be lost during these purification and wash steps. Finally, size selection will remove fragments outside of the specified size range required but some of the interacting fragments could also get removed.

The loss of material during the Hi-C experiment also makes the high number of starting cells a requirement. Rao *et al* generated Hi-C libraries from low cell numbers (2-5 million cells) (Rao *et al.* 2014) but in their protocol a 4-cutter restriction enzyme (*Mbo*I) was used which significantly increases library complexity and it would be possible to use more PCR cycles without creating duplicates. Fewer wash steps were also performed in the Rao experiments, potentially preserving yield. However, the amount of material they needed was considerably lower because they did not use a capture step. To sequence a Hi-C library only 20-50ng (Belton *et al.* 2012) is required, but a capture requires 400-750ng.

The amount of PCR amplification should generate enough library for sequencing without starting to make PCR duplicates. The yield from the biological replicate samples was not as high as the first samples, suggesting another round of PCR could possibly have been performed. Bioanalyzer analysis of pre-capture libraries determined that the average insert size for all the libraries was within acceptable ranges for Illumina sequencing. Therefore, as the Bioanalyzer assessment of the libraries showed the samples to be acceptable, the samples were still used for the capture experiments but with a lower input for the biological replicate samples. Post-capture QC using the

Bioanalyzer showed that all the samples, even those generated with a lower input, were suitable for sequencing.

6.4.3. Illumina sequencing and quality control

Hi-C libraries for NGS were produced using primers and adapters compatible with Illumina sequencing. Illumina sequencing has emerged as the industry leading standard. Early 'next generation sequencing' was pioneered by Roche, based on pyrosequencing and long reads of around 1000bp. The throughput of this technology, in the Mb of data, was superseded by technologies using short read lengths and bioinformatic tools to map back to a reference genome. Although other technologies are still available they have serious limitations when compared to the Illumina platform. The lead competitor, Ion Torrent, is again based on pyrosequencing, generating light when bases are added. This technology suffers in accuracy for longer, homogeneous sequences. Other technologies are emerging, such as PacBio and Oxford Nanopore, although neither of these is used routinely or robustly enough for general laboratory research. The Illumina platform also allows direct comparison of results obtained in different laboratories, where the published Hi-C protocols all specify Illumina sequencing for analysis. Illumina paired-end sequencing, with 50bp reads (Belton *et al.* 2012), identified large numbers of interacting fragments, with a short 50bp read length ensuring sequencing does not continue through the ligation junction into neighbouring fragments, making the mapping of interactions more accurate.

The flow cell for the MiSeq is a single lane, so the amount of sequencing data obtained is a maximum of 15Gb, 25 million reads, and 2x300bp read length. The HiSeq is a large-scale, high output instrument that can generate a maximum of 1000Gb of data, 4,000,000,000 reads, 2x125 read length in 2-11 days.

For the capture libraries, a full HiSeq lane would sequence all of the unique ligation products generated in the enriched libraries, and begin to sequence PCR artefacts creating duplicate sequences. Therefore, to maximise cost-efficiency, barcoding of samples was carried out during final library amplifications to allow multiple samples to be sequenced on the same lane, without any loss of full library capture. Final library amplifications were carried out using the minimum number of cycles, using multiple pooled reactions, to produce enough PCR product for sequencing, whilst minimising duplicate sequence artefacts.

Barcoded libraries were analysed using a MiSeq V3 150 cycle kit, for a final QC, before sending to the University of Manchester FLS Genomics Facility for full sequencing. The MiSeq did not possess the capacity to generate enough sequence data for an exhaustive interrogation of the Capture Hi-C (CHi-C) libraries, but as the technology is the same employed on the HiSeq it gave the best indication as to how the larger, expensive run will perform. Following the MiSeq run, QC of the completed libraries using HiCUP (discussed below) determined that the CHi-C libraries for the GM12878 and Jurkat cell lines were of sufficiently high quantity and quality to continue with full

HiSeq sequencing, based on the high percentage of valid and unique di-tags and the low percentage of background *trans* interactions.

Mapping and filtering of the CHi-C sequencing data was carried out in-house using HiCUP which is specifically designed for the analysis of Hi-C libraries (Wingett *et al.* 2015). The overall library quality was assessed by the amount of random inter-chromosomal *trans* interactions which occur through random collisions. A good quality library should have as low a *trans* interaction count as possible, seen in this context as 'background' interactions, ideally less than 50% (personal communication with Dr Stefan Schoenfelder, Babraham Institute, Cambridge, UK). The libraries I produced were of consistently very good quality with *trans* interactions making up <20% of the total interactions.

The amount of *trans* interactions in my libraries is comparable to the recent study by Nagano *et al*, comparing in-solution with in-nucleus ligation protocols (Nagano *et al.* 2015). The results obtained by Nagano *et al* showed that libraries made using the in-nucleus ligation protocol had 10-14% *trans,* compared to 26-65% *trans* in the in-solution libraries. The results from the Nagano study, and this study, provides further evidence that in-nucleus ligation produces more genuine interactions compared to the original in-solution protocol, leading to less experimental noise.

The capture efficiency was measured in terms of the ratio of on-target reads (reads containing a baited *Hin*dIII fragment) compared to the number of unique di-tags. In the region capture, half a lane of a HiSeq was used per sample for sequencing compared to the promoter capture, which used a full lane of the HiSeq in order to obtain the required sequencing depth. From the region capture experiment, 60.9 million (GM12878) and 54.9 million (Jurkat) di-tags were on-target with an average of 21,170 reads per *Hin*dIII restriction fragment, resulting in a capture efficiency of 62%. In the promoter capture experiment, 121 million (GM12878) and 115 million (Jurkat) unique di-tags were on target with an average of 21,448 per *Hin*dIII restriction fragment, resulting in a capture efficiency of 70%. Taken together, the amount of on-target reads from both captures was approximately 65%, which is in-line with recently reported data from Schoenfelder *et a*l who obtained a capture efficiency of 65-71% and >10-fold enrichment of read-pairs involving promoters compared to an un-baited Hi-C library (Schoenfelder *et al.* 2015).

The data in this study reinforces previous work showing that utilising a target capture step is an excellent way of increasing the resolution of Hi-C data, particularly when a large amount of sequencing is carried out to increase the sequencing depth. Hi-C alone, without a capture step, originally mapped only 0.26% of sequencing reads in a B-cell line (GM06990) (Lieberman-Aiden *et al.* 2009). A recent study by Mifsud *et al* (Mifsud *et al.* 2015) also demonstrated the low resolution of standard Hi-C in GM12878 B-cells, showing that only 45 million unique di-tags could be mapped and only 143 reads per restriction fragment were detected. This compared to approximately 21,000 reads per restriction fragment in my study, which is considerably higher.

Previous work by Dryden *et al* showed how using Capture Hi-C to study breast cancer susceptibility loci can significantly increase on-target reads (Dryden *et al.* 2014). Data obtained in the Dryden study was compared to a B-cell line, GM06990, used in the original Hi-C paper (Lieberman-Aiden *et al.* 2009). Without using a capture step only 2.3-5.9 million unique di-tags could be mapped, however, using a capture step resulted in 7.5-15% of reads being successfully mapped, which was a 30-60 fold increase over the original Hi-C protocol (Dryden *et al.* 2014; Lieberman-Aiden *et al.* 2009).

Further work investigating long-range interactions between promoters and regulatory elements by Mifsud *et al* obtained a 35-fold enrichment over non-targeted regions and 10-fold enrichment over a standard Hi-C with no capture (Mifsud *et al.* 2015). Jager *et al* obtained an average of 130-fold enrichment of target regions in a recent study of long-range interactions in colorectal cancer loci (Jager *et al.* 2015). Data from the Jager study was compared to the GM06990 Hi-C dataset (Lieberman-Aiden *et al.* 2009). From 30 million raw reads, 10-14 million reads were unique di-tags, of which 3.7-5.7 million reads were on-target (36-69%) compared to 28 thousand reads in the reference dataset giving an enrichment of 133-201 over the non-captured Hi-C data.

My work, therefore, compares extremely favourably with current published work. Indeed the average read depth I obtained, approximately 21,000 per targeted fragment, is the highest of any current study. This ensured that not only was my data of extremely high quality, it meant I could call interactions with a high degree of confidence. Interactions could be localised to a single restriction fragment and, importantly, found more long range interactions than in previous studies.

It is important to consider how quickly the protocols have progressed and improved since the development of 3C in 2002 and Hi-C in 2009. In 2014, Hughes et al used a variation of 3C called Capture C, which uses 3C followed by NGS, to interrogate hundreds of interactions simultaneously (Hughes *et al.* 2014). Although similar in design to Capture Hi-C, Capture C has several limitations. Since it is based on capturing a 3C library it lacks the crucial step of enriching for genuine ligation products. In Capture Hi-C biotin is incorporated into the ligation junction, such that a pull-down step can enrich for ligation products. By just performing a pull-down of the targeted sequence, as is performed in Capture C, all DNA containing this sequence is enriched, irrespective of whether it is in a ligation or not. This leads to the sequencing of a vast proportion of 'invalid' reads – as was seen in the Hughes paper. By incorporating Hi-C, performing ligation in the nucleus, and sequencing to a high depth, my capture Hi-C has vastly increased the resolution and application of this technique and made it a useful and efficient tool for post-GWAS analysis.

Summary

The Capture Hi-C summary statistics from this study show that I can reproducibly generate high quality CHi-C libraries, comparable to libraries produced by other groups in recent published studies. This will be a tremendous asset to the Manchester group and will enable new studies to

be undertaken, which will use CHi-C in primary immune cells under various conditions to further investigate disease associated variants and their target genes.

6.4.4. Capture Hi-C is a powerful tool to follow up on GWAS hits

Capture Hi-C implicates novel candidate genes in related autoimmune diseases

In the Capture Hi-C study, interactions were identified as true positives if they were seen in both biological replicates and in both capture experiments. Several false discovery thresholds were tested to give confidence that true interactions were being called. Increasing the stringency of the FDR threshold increased the enrichment of overlapping calls in the promoter and region capture experiments. Since the higher the FDR stringency the stronger evidence there is of an interaction, increasing the overlap gave me confidence the interactions called were real. This was confirmed using other available data. Publicly available datasets from Hi-C and 5C ENCODE experimental data were used to compare interaction data in similar cell lines, which confirmed interactions within the well-characterised *HBA* locus (Hughes *et al.* 2014) and with *IFNAR1* and *IL5* (Sanyal *et al.* 2012).

Interestingly, approximately 80% of the interactions occurred with non-promoter regions outside of the 500kb window and could not be validated within the design constraints of the complementary capture (promoters within 500kb of the index SNP). The GM12878 data was also compared to the Rao *et al* high-resolution Hi-C dataset (Rao *et al.* 2014), which validated 377 long-range (>500kb) interactions with an observed over expected ratio of >50, and confirmed long-range interactions with *FOXO1* and *ZFP36C1* which had been co-validated in the promoter and region capture experiments. In addition over 60% of the long range interactions found in my Capture Hi-C data were validated in the Rao Hi-C data at a observed over expected ratio of >10, giving reassurance that the findings in this study are real.

Interactions were identified between disease-associated regions and novel candidate genes which has given us more insight into the complexities of disease-associated loci and potentially identified some new causal genes which had not been previously considered in GWAS studies. One of the most interesting observations from our study was that regions containing loci for different autoimmune diseases, separated by a large distance, were shown to have common long-range interaction targets. For example, 16p13 SNPs associated independently with RA, psoriatic arthritis (PsA) and T1D could all interact with the *DEXI* promoter. RA associated variants located within an enhancer of *RAD51B* interacted with the promoter of *ZFP36L1*, which also contains SNPs associated with JIA. Variants associated with PsA, within *DENND1B* could also interact with *PTPRC*, a region independently associated with RA.

Interactions involving other RA loci, for example, *TNFAIP3* (discussed below), *STAT4* (Ji *et al.* 2010; Orozco *et al*. 2008; Remmers *et al*. 2007) and *ARID5B* (Eyre *et al.* 2012; Okada *et al.* 2012),

were also identified and showed that non-coding associated regions can 'skip' genes to interact with a number of more distant candidates. This backs up research showing that multiple genes can be influenced by the same promoter (Schoenfelder *et al.* 2010b; Schoenfelder *et al.* 2015), and shows a complicated relationship whereby enhancers containing causal variants can interact with the same promoter. Such a complicated relationship between enhancers and promoters shows that simply annotating a locus with the nearest plausible gene could be misleading.

Summary

Capture Hi-C in this study has provided interesting insights into how different autoimmune diseases can interact with common gene targets, and has implicated some new potential candidate genes within disease associated loci. This will greatly facilitate in the post-GWAS era in linking disease associated variants to their target genes.

Indeed, the importance of Capture Hi-C as a tool to follow up GWAS has been recognised. Blueprint is an international consortium (Adams *et al.* 2012; Martens *et al.* 2013) which has carried out whole genome promoter capture Hi-C on a wide range of primary cells, under various stimuli (data soon to be published). This will be an extremely valuable resource in the future.

6.5. Bioinformatic analysis of the 6q23 RA region

Bioinformatic analysis of the 6q23 RA associated region was carried out using a variety of databases to enable the prioritisation of SNPs for further investigation. A number of high profile, multi-centre initiatives are well underway, including ENCODE and Blueprint (ENCODE Project 2012; Martens *et al.* 2013; Rivera *et al.* 2013), to provide the scientific community with the genetic annotation required to guide the translation of GWAS findings. These initiatives provide publicly available data on all chromosomal regions, including chromatin accessibility (DNaseI-HS), potential regulatory activity (histone marks, transcription factor binding) and expression profiles (eQTL). The strength of these databases is the vast quantity of data that is being generated in different cell lines, although a weakness is the lack of connection between databases and even though there is a wealth of data, it is currently only generated on a limited number of cell lines under basal conditions. This makes specific functional work on stimulated, relevant primary cells imperative in order to fully link genotype to outcome.

Bioinformatic analysis of the lead RA associated SNP rs6920220 was carried out to identify SNPs with regulatory potential in order to identify the most plausible causal SNP. This region showed strong association to RA, the third most significant signal after HLA and *PTPN22* in the WTCCC GWAS (WTCCC 2007), with association confirmed in other RA GWAS (Eyre *et al.* 2012; Okada *et al.* 2014). Along with RA, the 6q23 region has also been associated with other autoimmune

diseases such as SLE (Graham *et al.* 2008), T1D, JIA, CeD (Trynka *et al.* 2009), ulcerative colitis (UC) and psoriasis (Vereecke *et al.* 2011).

Both an initial in-house fine mapping study and the Immunochip study supported the presence of several independently associated markers in 6q23. The strongest associated variant, rs6920220, mapped to an intergenic region between *TNFAIP3* and *OLIG3* (Orozco et al. 2009). However, due to LD, this SNP may not be the actual causal SNP within the region. Analysis of the region using Haploreg v4.1 showed that eight SNPs were highly correlated with the lead SNP rs6920220, including rs6927172 which was in complete LD (r^2 = 1). Analysis using RegulomeDB did not include data for two of the SNPs in LD, rs6933404 (r^2 = 0.89) and rs11757201 (r^2 = 1). This could be a potential source of bias in the results especially as one of the missing SNPs is in perfect LD with the lead SNP and could, therefore, also potentially be causal.

The rs6927172 SNP demonstrates a number of lines of evidence to support a function in disease causality, including mapping to an enhancer region in B-lymphoblastoid cell lines, primary stimulated Th17, and T_{REG} cells (ChromHMM chromatin state). It also maps to a region of open chromatin, characterised by DNaseI hypersensitivity, shows evidence of binding regulatory proteins and lies in a conserved region. Regulome DB also indicated that the rs6927172 SNP was most likely to have a functional effect. In support of these results, previous work within our department had demonstrated that rs6927172 had evidence of functionality. Luciferase reporter assays indicated that rs6927172 affected the regulatory activity of *TNFAIP3* transcription, and EMSAs indicated differential transcription factor binding to the different alleles of this variant, highlighting its plausibility as a candidate functional SNP (Elsby *et al.* 2010).

Further bioinformatic analysis of this region using Genevar (HapMap cell lines) and GTEx (whole blood) failed to highlight any eQTL evidence to either genes (*TNFAIP3*, *OLIG3*) or SNPs in the region. Therefore, no evidence of a direct correlation between genotype and gene expression has so far been uncovered in the tissue type or conditions studied in the publicly available datasets. This does not necessarily mean the associated SNPs are not influencing gene expression, given the right cell types and stimulatory conditions. Indeed, evidence is accumulating as to both cell-specific and stimulatory-specific response eQTLs (reQTLs) (Lee *et al.* 2014), genotype and transcription relationships (Westra *et al.* 2014; Wright *et al.* 2014).

A recent in-house analysis (manuscript in preparation) has suggested that the risk allele of the intergenic 6q23 variant rs6927172 correlates with increased expression of *IL20RA* in CD4+ T-cells. 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 (P= 0.02). Additionally, CD4+ T-cell whole genome expression data was available from a cohort of 102 early undifferentiated arthritis patients collected at baseline. 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 (P=

0.03). Whole genome expression data was also available in primary CD19+ B-cells but no eQTLs were found, further backing the evidence for cell-type specificity of gene expression.

No rs6927172 genotype-specific effects could be detected for *IFNGR1* expression levels in CD4+ and CD8+ T-cells. eQTLs, though, are context specific (Edwards *et al.* 2013; Maurano *et al.* 2012; Nica *et al.* 2010), and therefore, it would be interesting to explore whether the SNP influences *IFNGR1* expression in other cell types and/or under different stimulatory conditions.

Summary

Publicly available databases are extremely useful in directing post-GWAS prioritisation of SNPs for functional studies based on features such as enhancer regions, open chromatin and transcription factor binding. However, the data within them is only from a limited number of cell-types making it difficult to identify specific eQTLs. Identification of eQTLs within the 6q23 region, and complex regions in general, will ultimately require the interrogation of datasets from more cell-types when they become available.

6.6. Capture Hi-C in the 6q23 region

Disease associated variants at the chromosomal region 6q23 encompass a complex, non-coding genomic region containing enhancer elements and which lies some distance from the nearest gene. 6q23 is an important locus in autoimmunity and has been implicated in multiple diseases by GWAS, where independent variants have been found to be associated with different autoimmune diseases. To date, investigation of the functional consequences of disease associated variants have focussed almost exclusively on the most plausible causal gene within the locus, *TNFAIP3*.

TNFAIP3 is a good candidate gene due to having a well-known role in immunity. *TNFAIP3* has anti-inflammatory activity through the inhibition of NF-κB pathways. It is required for the termination of NF-κB inflammatory signals induced by TNF, CD-40, IL-1 and TLRs. *TNFAIP3* is strongly expressed in immune cells (BioGPS), giving further support to the role of *TNFAIP3* in immune processes, and the protein has been shown to be expressed in the synovium, the main active site of RA pathogenesis (Elsby *et al.* 2010).

Three linkage disequilibrium (LD) blocks containing disease variants reside within the 6q23 locus. One region, tagged by the rs6920220 SNP, contains SNPs associated with RA, SLE, coeliac disease (CeD), inflammatory bowel disease (IBD), psoriasis (Ps) and psoriatic arthritis (PsA). This LD block lies >180kb away from the nearest plausible gene, *TNFAIP3* (Coenen *et al.* 2009; Graham *et al.* 2008; Thomson *et al.* 2007; WTCCC 2007). A second region spanning approximately 100kb, tagged by rs7752903, has been associated with predisposition to RA, SLE and CeD and includes the *TNFAIP3* gene itself. There is evidence that a TT>A polymorphism located within this LD block, 42kb downstream of *TNFAIP3*, alters A20 (the protein encoded by *TNFAIP3*) expression through impaired delivery of NF-κB to the *TNFAIP3* promoter (Adrianto *et al.*

2011; Catrysse *et al.* 2014; Wang *et al.* 2013; Zhang *et al.* 2016). An additional association signal, tagged by rs610604, confers risk to Ps and PsA (Bowes *et al.* 2011; Zhang *et al.* 2015a).

Visualisation of the CHi-C data in B-cells and T-cells indicated that a complex pattern of long-range interactions were taking place within the 6q23 locus. Excitingly, many of these interactions involved novel potential candidate genes, such as *IL20RA* and *IFNGR1*, and lncRNAs. However, no interaction was detected that directly linked the LD block containing the lead rs6920220 SNP with *TNFAIP3*, predicted to be the causal gene within the region. The use of two complementary captures allowed some level of interaction validation, however some interactions from the region capture could not be validated because they involved an interaction with a gene promoter outside of the 500kb window. Therefore, long-range interactions in the 6q23 locus in the B-cell line were also validated by comparing the data to the largest GM12878 Hi-C dataset available (Rao *et al.* 2014) which showed a high degree of correlation giving confidence in the analysis and data.

In the 6q23 locus, the novel genes involved in long-range interactions with the RA associated region were *IL20RA* and *IFNGR1* which are both genes involved in autoimmunity. Both of these genes lie approximately 700kb from the associated region. The *IL20RA* gene is a member of the IL-10 family of cytokines encoding 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 (Pestka *et al.* 2004). Evidence suggests that this family of cytokines have a pro-inflammatory effect, and are essential in the activation of the epithelial innate immunity (Rutz *et al.* 2014). Expression of *IL20RA* has been detected in whole blood, T-cells, B-cells and monocytes (Su *et al.* 2004). 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 highly expressed in the synovium of RA patients, in local inflammatory sites (Hsu *et al.* 2015; Sakurai *et al.* 2008) 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 (Sakurai *et al.* 2008). IL-20 is also involved in angiogenesis by inducing endothelial cell proliferation and migration, and also up-regulates IL-6 and TNF-α (Hsu *et al.* 2015). Very interestingly, two recent clinical trials have demonstrated that anti-IL-20 monoclonal antibody is effective in the treatment of RA and psoriasis (Gottlieb *et al.* 2015; Hsu *et al.* 2016; Senolt *et al.* 2015).

The identification of interactions between RA SNPs and *IL20RA*, an existing drug target, shows that CHi-C is potentially a useful tool for identifying novel therapeutic targets or redirecting existing drugs. It has recently been suggested that selecting therapeutic targets with additional genetic data supporting its role could double the chance of the drug being successful in clinical improvement (Nelson *et al.* 2015).

The CHi-C experiment also suggested another potential novel causal gene in the 6q23 region, *IFNGR1*, which 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

induces the secretion of other pro-inflammatory cytokines (Hu *et al.* 2008b). It has been shown that an increased expression of *IFNGR1* in blood is associated with RA (Tang *et al.* 2015) and, coupled with the data from these experiments, could certainly be considered a potential RA causal gene.

Interactions with lncRNAs downstream of *TNFAIP3* may also be playing a role in gene regulation within the region. There are nine RA associated regions which overlap with lncRNAs, including the *TNFAIP3-OLIG3* region suggesting lncRNAs are important in disease susceptibility (Ding *et al.* 2015).

6.7. Validation of 6q23 Capture Hi-C interactions using 3C-qPCR

6.7.1. 3C controls and experimental design

Validation of the multiple interactions identified throughout the 6q23 locus was carried out using a targeted 3C approach which has been successfully used by several groups to identify long-range interactions. As the interaction targets within the locus were known, using 3C was a valid approach to use as a follow-up experiment.

Several approaches have been used for 3C assays. The first protocols used PCR followed by gel electrophoresis and quantification using gel analysis software, normalising to BAC control libraries (Dekker *et al.* 2002; Naumova *et al.* 2012). This approach is simple to perform but it lacks sensitivity, is error prone, and is only semi-quantitative TaqMan probes have also been used in 3C assays, which are sensitive and quantitative (Dryden *et al.* 2014; Hagege *et al.* 2007; Splinter *et al.* 2006). Splinter *et al* used TaqMan probes in 3C-qPCR assays to detect allele-specific effects in the β-globin gene by designing TaqMan probes targeting specific polymorphisms near restriction sites in the locus-control region of β-globin. The disadvantage of TaqMan probes is that they are expensive and if each interaction needs a different probe it would make the assay prohibitively expensive.

SYBR green qPCR is much less expensive than TaqMan and the assay is quantitative, however SYBR green is less specific as it amplifies all available dsDNA (Abou El Hassan *et al.* 2009) which precludes using a large amount of input meaning some less common interactions could be missed. However, SYBR green has successfully been used in 3C analysis for quantification (Comet *et al.* 2011) and by making use of melt curve analysis, reliable quantification of looping interactions can also be carried out (Abou El Hassan *et al.* 2009). For my 3C analysis, 3C-qPCR using SYBR green with a low amount of input was chosen, as recommended by collaborators at the Babraham Institute.

Primer design for the qPCR was very important so guidelines in Naumova *et al* (2012) were followed. A unidirectional format was used which oriented all primers in the same direction, on the same DNA strand. This ensured that only ligation products which had formed from head-to-head ligation would be detected, minimising the detection of uninformative ligation products. Primers

were designed using Primer 3 within 150bp of the restriction fragment, with similar distance in each primer to prevent amplification bias. Multiple primers for each fragment were designed and melt-curve analysis was carried out for each primer pair, discarding non-specific primers. Common problems in 3C PCR assays are poor amplification and non-specific products due to incomplete digestion, but these can be solved by optimising restriction digestion conditions or optimising the PCR conditions (Naumova *et al.* 2012).

For each interaction, primers were designed in the 'anchor' fragment, potential interacting fragment, non-interacting fragments and a short-range control for normalisation was designed within one or two restriction fragments of the anchor fragment. In 3C, the short-range interaction is much more likely to occur due to proximity ligation with the interaction frequency decreasing with distance. Therefore, the short-range control can be used to correct for different primer efficiencies within the experiment and acts as a positive control because the short-range interaction should give a strong interaction. The interaction with the test fragment was confirmed if the interaction frequency was higher than in the non-interacting control regions.

Digestion of an entire complex genome to make a control library would make it impossible to detect individual interactions by PCR. Early approaches used gel-purified PCR fragments spanning the restriction sites under investigation (Tolhuis *et al.* 2002). The DNA concentration was determined and equal amounts of each fragment were mixed, digested and ligated to create the control library. In the assay, the control library was mixed with digested and ligated genomic DNA to mimic the complexity of a 3C library. This protocol was simplified by the use of BAC clones (Palstra *et al.* 2003), who used multiple, minimally overlapping, BAC clones to interrogate the β-globin cluster.

If the region of interest is large, multiple BAC clones could be needed to span all the restriction fragments. Equimolar amounts of each BAC need to be digested and ligated in order to prevent bias and ensure that all possible interactions within the region can take place. Unfortunately, the company supplying the BAC clones are unable to confirm if they are correct so PCR validation is required. It was discovered through a test PCR that one of the original BAC clones chosen (RP11-771C9), which contained the lead RA SNP rs6920220 which we were most interested in, was incorrect meaning an alternative had to be obtained (CTD-2511N24). If alternative BACs are needed it could require more than one clone to cover the gap, potentially meaning a complete redesign of the BAC library. This is a significant limitation of the protocol which is difficult to overcome due to the complexities of the human genome.

6.7.2. 3C-qPCR can be used to validate CHi-C results

Initially, 3C-qPCR was used to confirm the statistically significant interactions in the 6q23 locus identified in CHi-C. Further analysis of interactions involving the LD block containing the lead intergenic RA SNP rs6920220, and closely correlated rs6927172 and rs35926684 SNPs in the adjacent *Hin*dIII fragment showed some interesting results. The interacting *Hin*dIII fragment

identified in the CHi-C analysis was within the boundaries of the SNPs LD block, but this fragment did not actually contain any SNPs directly associated with RA. Analysis of some of the adjacent *Hin*dIII fragments containing RA SNPs by 3C-qPCR actually identified a much stronger interaction in the fragment containing the rs6927172 and rs35926684 SNPs than in the fragment identified in the original CHi-C data. Therefore, whilst CHi-C identified a region containing a statistically significant interaction, the use of 3C in this case could further refine the location of the interaction making it an extremely useful tool for fine-mapping interactions.

Genotype specific 3C-qPCR showed increased interactions with the *IL20RA* gene in the presence of the risk allele of rs6927172 (G) compared with the non-risk allele. By contrast, the genotype-specific interaction was not observed for the rs6920220 variant. However, although bioinformatic evidence coupled with previous EMSA results (Elsby *et al.* 2010) suggests rs6927172 as the most likely causal SNP, 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.

Analysis of the CHi-C data suggested that certain interactions were cell-type specific. In the 3C-qPCR validation experiments each interaction was tested with the same GM12878 B-cell line and Jurkats tested in the CHi-C experiments. Interestingly, some of the interactions that had been detected in only one cell type by CHi-C could be identified in both cell lines in the 3C-qPCR experiments. For example, interactions between the SNPs LD block and lncRNAs were only detected in B-cells in the CHi-C analysis but were detected in both cell lines by 3C-qPCR. This could mean that the interaction is not limited to cell-type, or that the stringency of the CHi-C analysis resulted in the interaction not being called as statistically significant and therefore missed. An interaction was identified between *IL20RA* and *TNFAIP3* in both cell lines, however this was only a low level of interaction in 3C-qPCR and was not statistically significant.

In the 3C-qPCR analysis of the 6q23 region 3C libraries that I generated from primary synovial fibroblasts as part of a separate validation study were available for testing. Synovial fibroblasts, along with cells of the immune system, have been shown to be important in the establishment and progression of RA (Huber *et al.* 2006) so it was especially pertinent to have the opportunity to test these cells alongside the immune cell types. Interaction with *IL20RA* and *IFNGR1*, at the *Hin*dIII fragment containing the rs6927172 was detected in synovial fibroblasts. An interaction with the lncRNAs RP11-10J5.1 and RP11-240M16.1, with the same rs6927172 *Hin*dIII fragment, appeared to occur more frequently in the synovial fibroblasts compared to the B-cells and T-cells suggesting that these lncRNAs could be important in regulating synovial fibroblasts. Interestingly, no interaction could be detected between *IL20RA* and *TNFAIP3* in synovial fibroblasts whereas the interaction could be detected in B-cells and T-cells, although only at a low level. A significant interaction was detected between *IL20RA* and the lncRNA RP11-10J5.1 in synovial fibroblast which was not significant in the B-cells or T-cells, further suggesting that the lncRNAs in the 6q23 region have an important role in synovial fibroblast interactions. It would be very exciting to carry

out further, genotype-specific studies on synovial fibroblasts, however they are very difficult to obtain so this may not be possible.

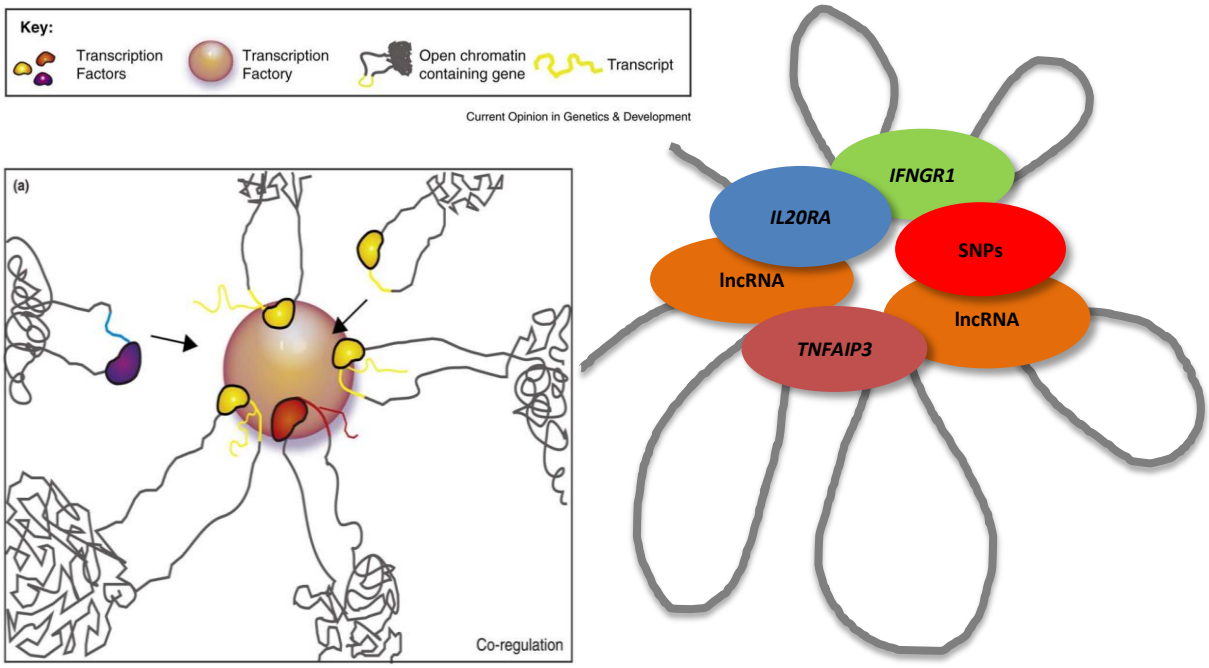
To assess if *IL20RA* interactions had cell-type specificity, in-house analysis of publicly available datasets performed by others in the group showed decreased or absent interactions in cell lines not expressing *IL20RA* such as HUVECs (human umbilical cord endothelial cells) and K562 (chronic myeloid leukaemia), providing further evidence for *IL20RA* as an important gene in autoimmunity.

Summary

Chromosome conformation capture experiments have revealed that the spatial organisation of the chromatin at the 6q23 region is complex, bringing together several genes with key roles in the immune response, including *IL20RA*, *IFNGR1* and *TNFAIP3*, along with regulatory elements containing SNPs associated with different autoimmune diseases (Figure 62 – adapted from Schoenfelder *et al.,* 2010). This supports the concept of transcription factories, where co-regulated genes come together to share transcription factors and regulatory elements such as enhancers (Schoenfelder *et al.* 2010b).

However, it cannot be ruled out that the interactions are occurring in a one-to-one manner, so if the enhancer is interacting with *IL20RA* it is unable to interact with *TNFAIP3* or *IFNGR1.* A mixed population of cells could have all possible interactions taking place, and therefore potentially be captured as seen in this experiment.

Figure 62: Multiple genes, SNPs and lncRNAs contribute to complex interplay in the 6q23 region



6.8. Investigation of regulatory protein binding in the 6q23 region

ChIP is frequently used to investigate DNA-Protein interactions occurring within the cell. It can be used to determine if a specific protein such as a transcription factor interacts with a particular genomic region. ChIP has already been successfully used to translate genetic findings in cancer studies (Pomerantz *et al.* 2009), to gain insight into transcriptional promoters and enhancers (Heintzman *et al.* 2007; Visel *et al.* 2009), and to investigate allele-specific interactions (Heintzman *et al.* 2007; Knight *et al.* 2003; Verlaan *et al.* 2009).

One of the main limitations of the ChIP assay can be the choice of antibody. For all the assays it was possible to obtain ChIP-grade antibodies which were of a higher concentration and validated for use in this type of assay.

Bioinformatic analysis predicted the binding of a number of transcription factors in the *TNFAIP3-OLIG3* region and showed that the RA associated SNPs lied in an enhancer region. In a recent study by Fahr *et al* it was suggested that around 60% of autoimmune disease risk variants mapped to enhancer regions (Farh *et al.* 2015). RA risk variants were found to localise at enhancers and super-enhancers regulating genes responsible for cell-specific effects and response to stimuli.

The rs6927172 variant was predicted through bioinformatics to alter the binding motif for eight transcription factors including BCL3 and NF-κB. BCL3 is a transcriptional co-activator that inhibits the nuclear translocation of the NF-κB p50 subunit in the cytoplasm and contributes to the regulation of transcription of NF-κB target genes in the nucleus (Bours *et al.* 1993; Carmody *et al.* 2007). The NF-κB family of transcription factors are important mediators of inflammatory signalling and regulate anti-apoptotic genes critical for cell survival (Monaco *et al.* 2004a; Monaco *et al.* 2004b).

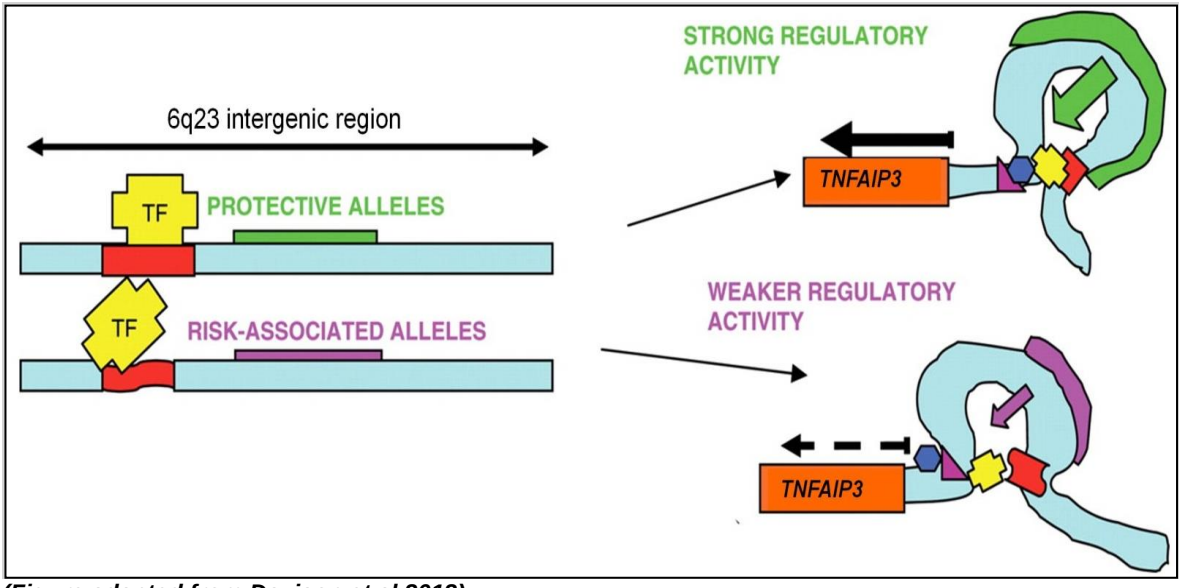
The NF-κB family of transcription factors contains five members: p50, p52, RelA (p65), c-Rel and RelB (Cui *et al.* 2014). All of the family members share a 300 amino acid Rel homology domain in the N-terminus which enables dimerisation. RelA, c-Rel and RelB contain activation domains located at the C-terminus which enable transcriptional activation of target genes. The p50 and p52 members are derived from larger precursors (p105/p100) and do not contain activation domains meaning they are transcriptional repressors unless they are bound to one of the other Rel subunits or BCL3.

The activity of NF-κB is inhibited by IκB protein family members until IκB is degraded by proteasomes, allowing the translocation of NF-κB to the nucleus where it can initiate gene expression. Activation of NF-κB leads to expression of *IκBα* which acts as a negative feedback loop. NF-κB can also be activated through toll-like receptors, which activate IκB kinases (Bonizzi *et al.* 2004; Ghosh *et al.* 1998; Moynagh 2005).

NF-κB is mostly known as an activator of transcription, but there is evidence that NF-kB family members can form homo- or heterodimers with most other family members, to produce gene regulatory complexes with different properties. For example, the p50–p50 dimer is a transcriptional repressor rather than activator. Also, NF-κB p65 can be an activator and repressor of its target genes depending upon the manner in which it is induced (Campbell *et al.* 2004).

Within the 6q23 region, NF-κB has the potential to bind to the rs6927172 protective allele, whereas the risk-associated allele could prevent or weaken transcription factor binding thereby affecting gene expression (summarised in Figure 63). For example, in the case of *TNFAIP3*, less expression due to reduced NF-κB signalling caused by the associated SNP could lead to increased inflammation, leading to the RA phenotype. *IL20RA* and *IFNGR1* are pro-inflammatory so more expression could lead to the RA phenotype. Therefore, the transcription factors chosen for investigation in this region were NF-κB and BCL3. The NF-κB members analysed were NF-κB1 (p105/p50), encoded by *NFKB1*, and RelA (p65), encoded by *RELA*.

Figure 63: The role of protective vs risk associated alleles in gene regulatory activity



(Figure adapted from Davison et al 2012)

RA SNPs in the 6q23 region bind transcription factors

Enrichment for the BCL3, NF-κB p50 and p65 transcription factors at the rs6927172 target region was detected in both B-cells and T-cells. However, statistical analysis of target region enrichment showed that there was no significant difference in binding between the genotypes in the B-cell lines. In Jurkat T-cells, however, TaqMan qPCR analysis performed by others in our research group using probes specific for each of the rs6927172 alleles has suggested a modest increase in NF-κB p65 enrichment in the presence of the risk allele (unpublished data, see Appendix Figure 71 - manuscript in preparation).

Interestingly, there were large differences in transcription factor enrichment between cell lines with the same genotype making it difficult to detect any genotype specific effects (see appendix Figures 68-70). Also, the number of cell lines available that were homozygous for the minor allele (GG) of rs6927172 was very small (only three cell lines available with this genotype) compared to the major allele homozygotes (CC = 10) or heterozygotes (CG = 8), which may also have affected any ability to detect if there were truly any genotype specific effects. The levels of enrichment between the two NF-κB subunits were different, with the p50 subunit more enriched than p65. This is not surprising, considering that p50 can form a heterodimer with BCL3 and both transcription factors were enriched at the target region.

Differential transcription factor binding has been previously demonstrated by our group (Elsby *et al.* 2010), however, as this study utilised EMSAs, the exact transcription factor could not be determined. Many transcription factors bind in the target region, therefore it is possible other transcription factors, or histone modifications, could be being affected by the different genotypes. Also, genotype-specific ChIP on multiple samples with the same genotype was only possible to be carried out in B-cell lines due to different genotypes being unavailable for the Jurkat T-cell line, and T-cells are thought to be more relevant for RA susceptibility.

The use of TaqMan assays in experiments recently performed by others in our research group has allowed some level of allele-specificity to be investigated in the T-cell line as they are heterozygous for the SNP, but it would be very useful to assess genotype effects in primary CD4+ T-cells isolated from individuals with the appropriate SNP genotype.

The failure to detect differential binding in B-cells may not be altogether surprising, given that ultimately this should correlate to an effect on expression levels and there is a lack of evidence for an eQTL in these cell lines in the region. It may well require a specific stimulation, or a different subset of cell type, to show a difference between the associated and non-associated genotype in terms of genetic function.

RA SNPs in the 6q23 region lie in an enhancer region

As previously discussed, many GWAS intergenic SNPs lie in regulatory regions of the genome and evidence from bioinformatics analysis showed that the rs6927172 SNP lies in an enhancer region. ChIP followed by SYBR green qPCR for the histone marks H3K4me1 and H3K27ac, signifying active enhancers, in both T-cells and B-cells suggested that the 6q23 SNPs did indeed lie in an enhancer region, reinforcing the evidence obtained from bioinformatics analysis.

Different HapMap B-cell lines containing the relevant rs6927172 genotypes were analysed using SYBR green qPCR which demonstrated a statistically significant increase in histone mark enrichment at the target region in samples containing the non-risk allele. As previously discussed, TaqMan assays have since been employed by others in our research group to test allele-specificity in the Jurkat T-cells as they are heterozygous (CG) for the rs6927172 target SNP. The data from these assays has suggested a modest increase in enhancer mark enrichment in the presence of the risk allele (unpublished data, see Appendix Figure 71 - manuscript in preparation). This data suggests that the risk allele could potentially have more of an effect on gene expression in T-cells. To follow on from these results it would be interesting to perform the same assay in primary T-cells from different individuals carrying the different SNP alleles.

Very interestingly, the risk variant of rs6927172 had opposite effects on enhancer mark enrichments in the cell lines tested. This highlights the cell specificity of gene regulation and it has been suggested that up to 50% of allele specific associations with epigenetic marks of enhancer activity (histone eQTLs) show an inconsistent direction of effect (Kilpinen *et al.* 2013).

Summary

Evidence obtained from Capture Hi-C, targeted 3C-qPCR and bioinformatics all suggested that the rs6927172 SNP, which is in perfect LD with the GWAS index SNP rs6920220, had a possible functional role.

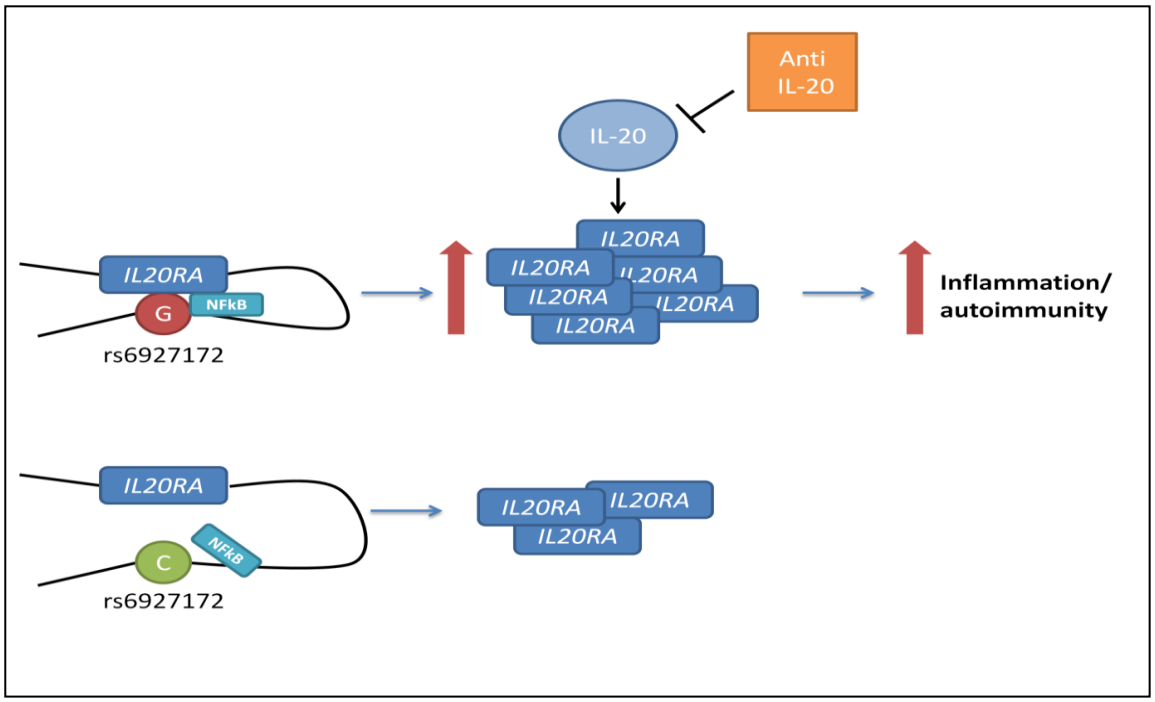
ChIP analysis demonstrated that transcription factors could bind to the SNP target region, the rs6927172 SNP lies in an enhancer region and that the risk allele could potentially have an effect on gene expression based on increased enrichment of enhancer marks in the presence of the risk allele.

A potential regulatory model for the 6q23 SNPs

Taken together, these results suggest that the mechanism by which the risk allele of rs6927172 alters expression of genes such as *IL20RA, IFNGR1* and *TNFAIP3* may be mediated by an increased regulatory activity and augmented binding of the NF-κB transcription factor.

This is summarised in the model below using *IL20RA* as an example (Figure 64). In the presence of the risk allele, NF-κB and *IL20RA* are recruited to the enhancer region containing the SNP resulting in upregulation of *IL20RA*, which binds IL-20, the target of an existing drug. This results in increased inflammation and the autoimmunity phenotype.

Figure 64: The rs6927172 risk allele (G) increases expression of *IL20RA* through increased regulatory activity and augmented binding of NF-κB



6.9. Conclusions

Here I present findings from a systematic approach to identify causal genes at the 6q23 RA locus using the recently developed CHi-C method. The use of CHi-C in this study has shown it to be an extremely valuable tool for following up on GWAS hits, identifying a novel candidate gene in the 6q23 locus which has independently been developed as a drug target.

The results from this study reinforces previous evidence that the nearest plausible biological candidate gene is not necessarily the causal gene. In the 6q23 locus, whilst *TNFAIP3* gene involvement is still implicated, another potential causal gene may well be *IL20RA*. Evidence that *IL20RA* could well be a target gene for RA is supported by the success of anti-IL20 therapies in RA and psoriasis.

This study illustrates the challenges in linking disease associated variants to function, showing that associated variants can be linked to a number of genes, dependent on which enhancer they are located within and the cell type under investigation.

6.10. Strengths and weaknesses of this project

The challenge in the post-GWAS era is to link disease associated variants to function. In this study, bioinformatic and molecular investigations were used to study RA loci and successfully identified potential novel candidate genes and functional SNPs in the 6q23 autoimmunity locus. In this section, the strengths and weaknesses of this project are outlined.

Strengths

The main strength in this project was the experimental lab work. The introduction of complex techniques such as Capture Hi-C into the laboratory has enabled the translation of RA GWAS findings into the identification of novel causal genes and causal variants within the 6q23 autoimmunity locus in this project. Post-GWAS laboratory investigations are imperative in order to fully understand the mechanisms of complex diseases, so having these techniques established in this department is now enabling other research staff and students to undertake new projects studying GWAS loci in other autoimmune diseases.

The quality and reproducibility of the Capture Hi-C libraries I generated was excellent and comparable to recently published data from other groups, giving confidence that the data generated was of a high standard.

Capture Hi-C was performed using two complementary experiments targeting gene promoters and disease associated regions identified through GWAS and fine-mapping studies. This type of approach had not been carried out on complex diseases before this study. The depth of sequencing enabled high resolution analysis, which has greatly improved on previously published studies using this technique.

3C-qPCR could be used to further refine interactions identified in the CHi-C experiments, making it a useful tool for interaction validation in selected genomic regions. Using chromosome conformation capture experiments in this study highlighted how complex the interactions within autoimmunity loci such as 6q23 can be, making these techniques ideal for studying GWAS loci.

Weaknesses

The main weakness of this study was that the work was performed in cell lines which are not a particularly good model for primary cells. Ideally primary cells important in RA pathogenesis, under various stimulatory conditions would have been used but it would have been difficult to obtain the cell numbers required for the experiments.

The region capture experiment identified a surprising number of long range interactions in excess of 1Mb which were unable to be validated using the complementary promoter capture experiment. This could have been improved by analysing the region capture first, then designing the promoter capture based on those results to allow more cross-validation of enhancer-promoter interactions.

The ChIP analysis was quite limited – only two histone marks and three transcription factors were chosen for analysis. The T-cell line showed opposite allele effects to the B-cells in the assays which is somewhat confusing, but this was only one cell line. It would be interesting to repeat the experiments on genotyped, primary cells, under stimulatory conditions to obtain a clearer picture.

There was limited linking of molecular biology, e.g. CHi-C interactions and ChIP, to cellular function in this study. Even though this study provided compelling evidence of a potential new RA candidate gene in the 6q23 region and identified the most likely functional SNP, the contribution to disease pathogenesis remains unclear. To fully understand the effects of SNPs in disease pathogenesis, it will be important to follow up the results from this work with more experiments to determine exactly what the mechanism of action is, for example, reporter assays in the appropriate cell types could be used to confirm differential enhancer activity according to genotype in the SNP region and relate this to gene expression.

The effect of directly perturbing the associated region was not investigated. In the 6q23 region, it would be exciting to see the effects of deletion/modification of the LD block containing the RA SNPs. This could be carried out through genome editing techniques such as CRISPR/Cas9 followed up by gene expression assays, and would provide evidence as to which is the exact causal gene in the region.

6.11. Future work

The results presented in this thesis show how novel molecular techniques can be used to follow up on GWAS signals, with the ability to refine associated regions and identify new genes.

Future work will address the limitations of this study:

1. Work was performed in cell lines which are not a particularly good model for primary cells
2. There was limited linking of molecular biology, e.g. Hi-C interactions and ChIP, to cellular function
3. The effect of directly perturbing the associated region was not investigated

These limitations will be addressed with three strategies:

1. Work will now be optimised for primary, CD4+ T-cells
2. Interactions will be measured and then the downstream consequences, in terms of gene expression, will be monitored in a time course experiment
3. Tools will be developed in the form of genome editing, and specifically CRISPR, to perform the array of functional assays in primary cells, then perturb the associated enhancer region, then re-perform the functional tests to gain insight into how the genetic associations increase risk of disease

A significant limitation of the study was the use of cell lines not primary immune cells. Cell lines were chosen because of the convenience and the ability to generate the large cell numbers needed for the development and optimisation of experiments. Primary cells would be more phenotypically representative of the immune cells *in-situ* than cell lines, however the high cell numbers needed would be considerably more difficult to obtain without some level of sample pooling.

In the publicly available datasets no eQTLs have been determined in the 6q23 region. It may well require stimulation of separated primary cells to discover differences in transcription factor binding/chromatin folding in disease-associated regions. Indeed, a suggestive correlation between the rs6927172 SNP and *IL20RA* in CD4+ T-cells from healthy individuals and early RA patients has been detected in an in-house eQTL study.

Work is now underway in the laboratory assessing chromatin interaction profiles, related to functional outcome in the form of regulated gene expression using primary CD4+ T-cells under stimulatory conditions over a time-course. This will involve isolating primary CD4+ T-cells from PBMCs, pooling samples and relating enhancer/promoter interactions to nascent RNA production using RNA-seq. This is work is likely to provide vital experimental evidence as to the interactions involving RA associated enhancers that result in significant functional changes in gene expression.

Further CHi-C experiments are also ongoing using CD4+ T-cells isolated from RA patients with low and high disease activity, constituting natural stimulatory conditions, such that any changes in DNA conformation that are related to a naturally active disease state can be identified.

In the future, the development of CRISPR/Cas9 genome editing tools will enable a targeted approach to investigate the effects of specific SNPs on target genes. CRISPR could provide definitive, empirical evidence that changing a SNP or haplotype has a measurable effect. Genome editing could be used to change a risk SNP to a non-risk SNP, or vice-versa, and measure the effect on gene expression or interactions with regulatory proteins. Recently, genome editing in Parkinson's disease identified a risk variant in a non-coding enhancer element that regulated a key gene involved in disease pathogenesis (Soldner *et al.* 2016).

In addition this type of technology can be employed to cleave out regions of the genome, for example TAD boundaries, to determine the downstream consequences. Interestingly, the RA SNP rs6927172 lies on a TAD boundary and it has been recently suggested that the perturbation of TAD boundaries is implicated in complex diseases (Lupianez *et al.* 2015) and in cancer (Hnisz *et al.* 2016; Valton *et al.* 2016). The role of the SNP in TAD arrangements will certainly be worth further investigation in the future.

Several recent studies have used genome editing techniques to delete regulatory elements and identify target genes, for example Li et al deleted a section of a downstream *SOX2* gene enhancer resulting in a marked decrease in gene expression (Li *et al.* 2014). Claussnitzer *et al* (Claussnitzer *et al.* 2014) have used CRISPR/Cas9 and other tools to investigate a type-2-diabetes associated variant in the *PPAGR2* gene and showed that replacing the risk allele with the non-risk allele could increase expression of the transcript. The same group also investigated a FTO variant strongly associated with obesity (Claussnitzer *et al.* 2015) and used CRISPR/Cas9 editing to repair a conserved motif.

Alternatively, dead cas9 (dCas9) which lacks nuclease activity and does not cut the DNA could be employed to deliver either stimulatory or inhibitory molecules to the site of the implicated enhancers. This could then be used to investigate the downstream consequences of the enhancer, and which gene it regulates (Gao *et al.* 2014; Pham *et al.* 2016). In future these dCas9 regulatory systems can be used as a multiplexed system to up and downregulate enhancers on pathways implicated to play a key role in disease by genetic analysis.

Evidence of cell-type specific interactions means that this study is likely to be only the beginning of similar explorations. Further work to characterise functionally the observed interactions are required to determine how disease associated SNPs influence the risk of disease, with the aim of better understanding disease aetiology.

7. References

References

Abecasis, G. R., Auton, A., Brooks, L. D., DePristo, M. A., Durbin, R. M., Handsaker, R. E. *et al.* (2012). An integrated map of genetic variation from 1,092 human genomes. *Nature*, **491**, (7422), 56-65.

Abou El Hassan, M. and Bremner, R. (2009). A rapid simple approach to quantify chromosome conformation capture. *Nucleic Acids Res*, **37**, (5), e35.

Abraham, R. T. and Weiss, A. (2004). Jurkat T cells and development of the T-cell receptor signalling paradigm. *Nat Rev Immunol*, **4**, (4), 301-308.

Adams, D., Altucci, L., Antonarakis, S. E., Ballesteros, J., Beck, S., Bird, A. *et al.* (2012). BLUEPRINT to decode the epigenetic signature written in blood. *Nat Biotechnol*, **30**, (3), 224-226.

Adli, M. and Bernstein, B. E. (2011). Whole-genome chromatin profiling from limited numbers of cells using nano-ChIP-seq. *Nat Protoc*, **6**, (10), 1656-1668.

Adrianto, I., Wen, F., Templeton, A., Wiley, G., King, J. B., Lessard, C. J. *et al.* (2011). Association of a functional variant downstream of TNFAIP3 with systemic lupus erythematosus. *Nat Genet*, **43**, (3), 253-258.

Ahmadiyeh, N., Pomerantz, M. M., Grisanzio, C., Herman, P., Jia, L., Almendro, V. *et al.* (2010). 8q24 prostate, breast, and colon cancer risk loci show tissue-specific long-range interaction with MYC. *Proc Natl Acad Sci U S A*, **107**, (21), 9742-9746.

Andersson, R., Gebhard, C., Miguel-Escalada, I., Hoof, I., Bornholdt, J., Boyd, M. *et al.* (2014). An atlas of active enhancers across human cell types and tissues. *Nature*, **507**, (7493), 455-461.

Apostolou, E., Ferrari, F., Walsh, R. M., Bar-Nur, O., Stadtfeld, M., Cheloufi, S. *et al.* (2013). Genome-wide chromatin interactions of the Nanog locus in pluripotency, differentiation, and reprogramming. *Cell Stem Cell*, **12**, (6), 699-712.

Apostolou, E. and Thanos, D. (2008). Virus Infection Induces NF-kappaB-dependent interchromosomal associations mediating monoallelic IFN-beta gene expression. *Cell*, **134**, (1), 85-96.

Arand, J., Spieler, D., Karius, T., Branco, M. R., Meilinger, D., Meissner, A. *et al.* (2012). In vivo control of CpG and non-CpG DNA methylation by DNA methyltransferases. *PLoS Genet*, **8**, (6), e1002750.

Ban, M., Goris, A., Lorentzen, A. R., Baker, A., Mihalova, T., Ingram, G. *et al.* (2009). Replication analysis identifies TYK2 as a multiple sclerosis susceptibility factor. *Eur J Hum Genet*, **17**, (10), 1309-1313.

Bantignies, F. and Cavalli, G. (2006). Cellular memory and dynamic regulation of polycomb group proteins. *Curr Opin Cell Biol*, **18**, (3), 275-283.

Bartel, D. P. (2009). MicroRNAs: target recognition and regulatory functions. *Cell*, **136**, (2), 215-233.

Bauer, D. E., Kamran, S. C., Lessard, S., Xu, J., Fujiwara, Y., Lin, C. *et al.* (2013). An erythroid enhancer of BCL11A subject to genetic variation determines fetal hemoglobin level. *Science*, **342**, (6155), 253-257.

Begovich, A. B., Carlton, V. E., Honigberg, L. A., Schrodi, S. J., Chokkalingam, A. P., Alexander, H. C. *et al.* (2004). A missense single-nucleotide polymorphism in a gene encoding a protein tyrosine phosphatase (PTPN22) is associated with rheumatoid arthritis. *Am J Hum Genet*, **75**, (2), 330-337.

Belkaid, Y. and Hand, T. W. (2014). Role of the microbiota in immunity and inflammation. *Cell*, **157**, (1), 121-141.

Belton, J. M., McCord, R. P., Gibcus, J. H., Naumova, N., Zhan, Y., and Dekker, J. (2012). Hi-C: a comprehensive technique to capture the conformation of genomes. *Methods*, **58**, (3), 268-276.

Bernstein, B. E., Stamatoyannopoulos, J. A., Costello, J. F., Ren, B., Milosavljevic, A., Meissner, A. *et al.* (2010). The NIH Roadmap Epigenomics Mapping Consortium. *Nat Biotechnol*, **28**, (10), 1045-1048.

Bertone, P., Gerstein, M., and Snyder, M. (2005). Applications of DNA tiling arrays to experimental genome annotation and regulatory pathway discovery. *Chromosome Res*, **13**, (3), 259-274.

Bertucci, F., Lagarde, A., Ferrari, A., Finetti, P., Charafe-Jauffret, E., Van, L. S. *et al.* (2012). 8q24 Cancer risk allele associated with major metastatic risk in inflammatory breast cancer. *PLoS One*, **7**, (5), e37943.

Bira, Y., Tani, K., Nishioka, Y., Miyata, J., Sato, K., Hayashi, A. *et al.* (2005). Transforming growth factor beta stimulates rheumatoid synovial fibroblasts via the type II receptor. *Mod Rheumatol*, **15**, (2), 108-113.

Bolger, A. M., Lohse, M., and Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*, **30**, (15), 2114-2120.

Bonizzi, G. and Karin, M. (2004). The two NF-kappaB activation pathways and their role in innate and adaptive immunity. *Trends Immunol*, **25**, (6), 280-288.

Bottini, N. and Firestein, G. S. (2013). Epigenetics in rheumatoid arthritis: a primer for rheumatologists. *Curr Rheumatol Rep*, **15**, (11), 372.

Bottini, N., Musumeci, L., Alonso, A., Rahmouni, S., Nika, K., Rostamkhani, M. *et al.* (2004). A functional variant of lymphoid tyrosine phosphatase is associated with type I diabetes. *Nat Genet*, **36**, (4), 337-338.

Bours, V., Franzoso, G., Azarenko, V., Park, S., Kanno, T., Brown, K. *et al.* (1993). The oncoprotein Bcl-3 directly transactivates through kappa B motifs via association with DNA-binding p50B homodimers. *Cell*, **72**, (5), 729-739.

Bowes, J., Budu-Aggrey, A., Huffmeier, U., Uebe, S., Steel, K., Hebert, H. L. *et al.* (2015). Dense genotyping of immune-related susceptibility loci reveals new insights into the genetics of psoriatic arthritis. *Nat Commun*, **6**, 6046.

Bowes, J., Orozco, G., Flynn, E., Ho, P., Brier, R., Marzo-Ortega, H. *et al.* (2011). Confirmation of TNIP1 and IL23A as susceptibility loci for psoriatic arthritis. *Ann Rheum Dis*, **70**, (9), 1641-1644.

Boyle, A. P., Hong, E. L., Hariharan, M., Cheng, Y., Schaub, M. A., Kasowski, M. *et al.* (2012). Annotation of functional variation in personal genomes using RegulomeDB. *Genome Res*, **22**, (9), 1790-1797.

Brentano, F., Kyburz, D., and Gay, S. (2009). Toll-like receptors and rheumatoid arthritis. *Methods Mol Biol*, **517**, 329-343.

Bulger, M. and Groudine, M. (2011). Functional and mechanistic diversity of distal transcription enhancers. *Cell*, **144**, (3), 327-339.

Bulik-Sullivan, B., Selitsky, S., and Sethupathy, P. (2013). Prioritization of genetic variants in the microRNA regulome as functional candidates in genome-wide association studies. *Hum Mutat*, **34**, (8), 1049-1056.

Butter, F., Davison, L., Viturawong, T., Scheibe, M., Vermeulen, M., Todd, J. A. *et al.* (2012). Proteome-wide analysis of disease-associated SNPs that show allele-specific transcription factor binding. *PLoS Genet*, **8**, (9), e1002982.

Campbell, K. J., Rocha, S., and Perkins, N. D. (2004). Active repression of antiapoptotic gene expression by RelA(p65) NF-kappa B. *Mol Cell*, **13**, (6), 853-865.

Carmody, R. J. and Chen, Y. H. (2007). Nuclear factor-kappaB: activation and regulation during toll-like receptor signaling. *Cell Mol Immunol*, **4**, (1), 31-41.

Carrel, L. and Willard, H. F. (2005). X-inactivation profile reveals extensive variability in X-linked gene expression in females. *Nature*, **434**, (7031), 400-404.

Catrysse, L., Vereecke, L., Beyaert, R., and van, L. G. (2014). A20 in inflammation and autoimmunity. *Trends Immunol*, **35**, (1), 22-31.

Cheung, V. G., Spielman, R. S., Ewens, K. G., Weber, T. M., Morley, M., and Burdick, J. T. (2005). Mapping determinants of human gene expression by regional and genome-wide association. *Nature*, **437**, (7063), 1365-1369.

Choy, E. (2008). Inhibiting interleukin-6 in rheumatoid arthritis. *Curr Rheumatol Rep*, **10**, (5), 413-417.

Choy, E. (2012). Understanding the dynamics: pathways involved in the pathogenesis of rheumatoid arthritis. *Rheumatology (Oxford)*, **51 Suppl 5**, v3-11.

Choy, E. H., Kavanaugh, A. F., and Jones, S. A. (2013). The problem of choice: current biologic agents and future prospects in RA. *Nat Rev Rheumatol*, **9**, (3), 154-163.

Christova, R. (2013). Detecting DNA-protein interactions in living cells-ChIP approach. *Adv Protein Chem Struct Biol*, **91**, 101-133.

Claussnitzer, M., Dankel, S. N., Kim, K. H., Quon, G., Meuleman, W., Haugen, C. *et al.* (2015). FTO Obesity Variant Circuitry and Adipocyte Browning in Humans. *N Engl J Med*, **373**, (10), 895-907.

Claussnitzer, M., Dankel, S. N., Klocke, B., Grallert, H., Glunk, V., Berulava, T. *et al.* (2014). Leveraging cross-species transcription factor binding site patterns: from diabetes risk loci to disease mechanisms. *Cell*, **156**, (1-2), 343-358.

Clop, A., Marcq, F., Takeda, H., Pirottin, D., Tordoir, X., Bibe, B. *et al.* (2006). A mutation creating a potential illegitimate microRNA target site in the myostatin gene affects muscularity in sheep. *Nat Genet*, **38**, (7), 813-818.

Coenen, M. J. and Gregersen, P. K. (2009). Rheumatoid arthritis: a view of the current genetic landscape. *Genes Immun*, **10**, (2), 101-111.

Collas, P. (2010). The current state of chromatin immunoprecipitation. *Mol Biotechnol*, **45**, (1), 87-100.

Collas, P. and Dahl, J. A. (2008). Chop it, ChIP it, check it: the current status of chromatin immunoprecipitation. *Front Biosci*, **13**, 929-943.

Comet, I., Schuettengruber, B., Sexton, T., and Cavalli, G. (2011). A chromatin insulator driving three-dimensional Polycomb response element (PRE) contacts and Polycomb association with the chromatin fiber. *Proc Natl Acad Sci U S A*, **108**, (6), 2294-2299.

Corradin, O. and Scacheri, P. C. (2014). Enhancer variants: evaluating functions in common disease. *Genome Med*, **6**, (10), 85.

Couturier, N., Bucciarelli, F., Nurtdinov, R. N., Debouverie, M., Lebrun-Frenay, C., Defer, G. *et al.* (2011). Tyrosine kinase 2 variant influences T lymphocyte polarization and multiple sclerosis susceptibility. *Brain*, **134**, (Pt 3), 693-703.

Cowper-Sal.lari R., Zhang, X., Wright, J. B., Bailey, S. D., Cole, M. D., Eeckhoute, J. *et al.* (2012). Breast cancer risk-associated SNPs modulate the affinity of chromatin for FOXA1 and alter gene expression. *Nat Genet*, **44**, (11), 1191-1198.

Cremer, T. and Cremer, C. (2006a). Rise, fall and resurrection of chromosome territories: a historical perspective. Part I. The rise of chromosome territories. *Eur J Histochem*, **50**, (3), 161-176.

Cremer, T. and Cremer, M. (2010). Chromosome territories. *Cold Spring Harb Perspect Biol*, **2**, (3), a003889.

Cremer, T., Cremer, M., Dietzel, S., Muller, S., Solovei, I., and Fakan, S. (2006b). Chromosome territories--a functional nuclear landscape. *Curr Opin Cell Biol*, **18**, (3), 307-316.

Cribbs, A. P., Kennedy, A., Penn, H., Read, J. E., Amjadi, P., Green, P. *et al.* (2014). Treg cell function in rheumatoid arthritis is compromised by ctla-4 promoter methylation resulting in a failure to activate the indoleamine 2,3-dioxygenase pathway. *Arthritis Rheumatol*, **66**, (9), 2344-2354.

Cui, J., Chen, Y., Wang, H. Y., and Wang, R. F. (2014). Mechanisms and pathways of innate immune activation and regulation in health and cancer. *Hum Vaccin Immunother*, **10**, (11), 3270-3285.

Cutolo, M. (2007). Sex and rheumatoid arthritis: mouse model versus human disease. *Arthritis Rheum*, **56**, (1), 1-3.

Cutolo, M., Capellino, S., Sulli, A., Serioli, B., Secchi, M. E., Villaggio, B. *et al.* (2006). Estrogens and autoimmune diseases. *Ann N Y Acad Sci*, **1089**, 538-547.

Dahl, J. A. and Collas, P. (2008). MicroChIP--a rapid micro chromatin immunoprecipitation assay for small cell samples and biopsies. *Nucleic Acids Res*, **36**, (3), e15.

Davison, L. J., Wallace, C., Cooper, J. D., Cope, N. F., Wilson, N. K., Smyth, D. J. *et al.* (2012). Long-range DNA looping and gene expression analyses identify DEXI as an autoimmune disease candidate gene. *Hum Mol Genet*, **21**, (2), 322-333.

de Laat, W. and Dekker, J. (2012). 3C-based technologies to study the shape of the genome. *Methods*, **58**, (3), 189-191.

De Rycke, L., Peene, I., Hoffman, I. E., Kruithof, E., Union, A., Meheus, L. *et al.* (2004). Rheumatoid factor and anticitrullinated protein antibodies in rheumatoid arthritis: diagnostic value, associations with radiological progression rate, and extra-articular manifestations. *Ann Rheum Dis*, **63**, (12), 1587-1593.

de Wit, E. and de Laat, W. (2012). A decade of 3C technologies: insights into nuclear organization. *Genes Dev*, **26**, (1), 11-24.

DeAngelis, M. M., Wang, D. G., and Hawkins, T. L. (1995). Solid-phase reversible immobilization for the isolation of PCR products. *Nucleic Acids Res*, **23**, (22), 4742-4743.

Dechancie, J. and Houk, K. N. (2007). The origins of femtomolar protein-ligand binding: hydrogen-bond cooperativity and desolvation energetics in the biotin-(strept)avidin binding site. *J Am Chem Soc*, **129**, (17), 5419-5429.

Degner, J. F., Pai, A. A., Pique-Regi, R., Veyrieras, J. B., Gaffney, D. J., Pickrell, J. K. *et al.* (2012). DNase I sensitivity QTLs are a major determinant of human expression variation. *Nature*, **482**, (7385), 390-394.

Deighton, C. M., Walker, D. J., Griffiths, I. D., and Roberts, D. F. (1989). The contribution of HLA to rheumatoid arthritis. *Clin Genet*, **36**, (3), 178-182.

Dekker, J. (2006). The three 'C' s of chromosome conformation capture: controls, controls, controls. *Nat Methods*, **3**, (1), 17-21.

Dekker, J., Rippe, K., Dekker, M., and Kleckner, N. (2002). Capturing chromosome conformation. *Science*, **295**, (5558), 1306-1311.

Dermitzakis, E. T., Reymond, A., and Antonarakis, S. E. (2005). Conserved non-genic sequences - an unexpected feature of mammalian genomes. *Nat Rev Genet*, **6**, (2), 151-157.

Dey, B., Thukral, S., Krishnan, S., Chakrobarty, M., Gupta, S., Manghani, C. *et al.* (2012). DNA-protein interactions: methods for detection and analysis. *Mol Cell Biochem*, **365**, (1-2), 279-299.

Dimas, A. S., Deutsch, S., Stranger, B. E., Montgomery, S. B., Borel, C., Attar-Cohen, H. *et al.* (2009). Common regulatory variation impacts gene expression in a cell type-dependent manner. *Science*, **325**, (5945), 1246-1250.

Ding, J., Eyre, S., and Worthington, J. (2015). Genetics of RA susceptibility, what comes next? *RMD Open*, **1**, (1), e000028.

Diogo, D., Okada, Y., and Plenge, R. M. (2014). Genome-wide association studies to advance our understanding of critical cell types and pathways in rheumatoid arthritis: recent findings and challenges. *Curr Opin Rheumatol*, **26**, (1), 85-92.

Dixon, J. R., Selvaraj, S., Yue, F., Kim, A., Li, Y., Shen, Y. *et al.* (2012). Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature*, **485**, (7398), 376-380.

Djebali, S., Davis, C. A., Merkel, A., Dobin, A., Lassmann, T., Mortazavi, A. *et al.* (2012). Landscape of transcription in human cells. *Nature*, **489**, (7414), 101-108.

Dostie, J. and Bickmore, W. A. (2012). Chromosome organization in the nucleus - charting new territory across the Hi-Cs. *Curr Opin Genet Dev*, **22**, (2), 125-131.

Dostie, J. and Dekker, J. (2007a). Mapping networks of physical interactions between genomic elements using 5C technology. *Nat Protoc*, **2**, (4), 988-1002.

Dostie, J., Richmond, T. A., Arnaout, R. A., Selzer, R. R., Lee, W. L., Honan, T. A. *et al.* (2006). Chromosome Conformation Capture Carbon Copy (5C): a massively parallel solution for mapping interactions between genomic elements. *Genome Res*, **16**, (10), 1299-1309.

Dostie, J., Zhan, Y., and Dekker, J. (2007b). Chromosome conformation capture carbon copy technology. *Curr Protoc Mol Biol*, **Chapter 21**, Unit.

Dryden, N. H., Broome, L. R., Dudbridge, F., Johnson, N., Orr, N., Schoenfelder, S. *et al.* (2014). Unbiased analysis of potential targets of breast cancer susceptibility loci by Capture Hi-C. *Genome Res*, **24**, (11), 1854-1868.

Duan, Z., Andronescu, M., Schutz, K., Lee, C., Shendure, J., Fields, S. *et al.* (2012). A genome-wide 3C-method for characterizing the three-dimensional architectures of genomes. *Methods*, **58**, (3), 277-288.

Duggal, G., Wang, H., and Kingsford, C. (2014). Higher-order chromatin domains link eQTLs with the expression of far-away genes. *Nucleic Acids Res*, **42**, (1), 87-96.

Edgar, R., Domrachev, M., and Lash, A. E. (2002). Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res*, **30**, (1), 207-210.

Edwards, C. J. (2008). Commensal gut bacteria and the etiopathogenesis of rheumatoid arthritis. *J Rheumatol*, **35**, (8), 1477-14797.

Edwards, S. L., Beesley, J., French, J. D., and Dunning, A. M. (2013). Beyond GWASs: illuminating the dark road from association to function. *Am J Hum Genet*, **93**, (5), 779-797.

Elsby, L. M., Orozco, G., Denton, J., Worthington, J., Ray, D. W., and Donn, R. P. (2010). Functional evaluation of TNFAIP3 (A20) in rheumatoid arthritis. *Clin Exp Rheumatol*, **28**, (5), 708-714.

ENCODE Project (2012). An integrated encyclopedia of DNA elements in the human genome. *Nature*, **489**, (7414), 57-74.

Ethier, S. D., Miura, H., and Dostie, J. (2012). Discovering genome regulation with 3C and 3C-related technologies. *Biochim Biophys Acta*, **1819**, (5), 401-410.

Eyre, S., Bowes, J., Diogo, D., Lee, A., Barton, A., Martin, P. *et al.* (2012). High-density genetic mapping identifies new susceptibility loci for rheumatoid arthritis. *Nat Genet*, **44**, (12), 1336-1340.

Fairfax, B. P., Humburg, P., Makino, S., Naranbhai, V., Wong, D., Lau, E. *et al.* (2014). Innate immune activity conditions the effect of regulatory variants upon monocyte gene expression. *Science*, **343**, (6175), 1246949.

Farh, K. K., Marson, A., Zhu, J., Kleinewietfeld, M., Housley, W. J., Beik, S. *et al.* (2015). Genetic and epigenetic fine mapping of causal autoimmune disease variants. *Nature*, **518**, (7539), 337-343.

Ferraiuolo, M. A., Sanyal, A., Naumova, N., Dekker, J., and Dostie, J. (2012). From cells to chromatin: capturing snapshots of genome organization with 5C technology. *Methods*, **58**, (3), 255-267.

Fitzgerald, K. A. and Caffrey, D. R. (2014). Long noncoding RNAs in innate and adaptive immunity. *Curr Opin Immunol*, **26**, 140-146.

Forrest, A. R., Kawaji, H., Rehli, M., Baillie, J. K., de Hoon, M. J., Haberle, V. *et al.* (2014). A promoter-level mammalian expression atlas. *Nature*, **507**, (7493), 462-470.

Fousteri, G., Liossis, S. N., and Battaglia, M. (2013). Roles of the protein tyrosine phosphatase PTPN22 in immunity and autoimmunity. *Clin Immunol*, **149**, (3), 556-565.

Fransen, J. and van Riel, P. L. (2006). DAS remission cut points. *Clin Exp Rheumatol*, **24**, (6 Suppl 43), S-32.

Franz, J. K., Kolb, S. A., Hummel, K. M., Lahrtz, F., Neidhart, M., Aicher, W. K. *et al.* (1998). Interleukin-16, produced by synovial fibroblasts, mediates chemoattraction for CD4+ T lymphocytes in rheumatoid arthritis. *Eur J Immunol*, **28**, (9), 2661-2671.

Fraser, P. and Bickmore, W. (2007). Nuclear organization of the genome and the potential for gene regulation. *Nature*, **447**, (7143), 413-417.

Freedman, M. L., Monteiro, A. N., Gayther, S. A., Coetzee, G. A., Risch, A., Plass, C. *et al.* (2011). Principles for the post-GWAS functional characterization of cancer risk loci. *Nat Genet*, **43**, (6), 513-518.

French, J. D., Ghoussaini, M., Edwards, S. L., Meyer, K. B., Michailidou, K., Ahmed, S. *et al.* (2013). Functional variants at the 11q13 risk locus for breast cancer regulate cyclin D1 expression through long-range enhancers. *Am J Hum Genet*, **92**, (4), 489-503.

Fu, J., Wolfs, M. G., Deelen, P., Westra, H. J., Fehrmann, R. S., Te Meerman, G. J. *et al.* (2012). Unraveling the regulatory mechanisms underlying tissue-dependent genetic variation of gene expression. *PLoS Genet*, **8**, (1), e1002431.

Fuks, F., Burgers, W. A., Brehm, A., Hughes-Davies, L., and Kouzarides, T. (2000). DNA methyltransferase Dnmt1 associates with histone deacetylase activity. *Nat Genet*, **24**, (1), 88-91.

Fulci, V., Scappucci, G., Sebastiani, G. D., Giannitti, C., Franceschini, D., Meloni, F. *et al.* (2010). miR-223 is overexpressed in T-lymphocytes of patients affected by rheumatoid arthritis. *Hum Immunol*, **71**, (2), 206-211.

Fullwood, M. J., Han, Y., Wei, C. L., Ruan, X., and Ruan, Y. (2010). Chromatin interaction analysis using paired-end tag sequencing. *Curr Protoc Mol Biol*, **Chapter 21**, Unit-25.

Fullwood, M. J., Liu, M. H., Pan, Y. F., Liu, J., Xu, H., Mohamed, Y. B. *et al.* (2009a). An oestrogen-receptor-alpha-bound human chromatin interactome. *Nature*, **462**, (7269), 58-64.

Fullwood, M. J. and Ruan, Y. (2009b). ChIP-based methods for the identification of long-range chromatin interactions. *J Cell Biochem*, **107**, (1), 30-39.

Gabriel, S. E. and Crowson, C. S. (2012). Risk factors for cardiovascular disease in rheumatoid arthritis. *Curr Opin Rheumatol*, **24**, (2), 171-176.

Gao, X., Tsang, J. C., Gaba, F., Wu, D., Lu, L., and Liu, P. (2014). Comparison of TALE designer transcription factors and the CRISPR/dCas9 in regulation of gene expression by targeting enhancers. *Nucleic Acids Res*, **42**, (20), e155.

Gavrilov, A. A., Gushchanskaya, E. S., Strelkova, O., Zhironkina, O., Kireev, I. I., Iarovaia, O. V. *et al.* (2013). Disclosure of a structural milieu for the proximity ligation reveals the elusive nature of an active chromatin hub. *Nucleic Acids Res*, **41**, (6), 3563-3575.

Geiler, J., Buch, M., and McDermott, M. F. (2011). Anti-TNF treatment in rheumatoid arthritis. *Curr Pharm Des*, **17**, (29), 3141-3154.

Georges, M. (2011). The long and winding road from correlation to causation. *Nat Genet*, **43**, (3), 180-181.

Ghosh, S., May, M. J., and Kopp, E. B. (1998). NF-kappa B and Rel proteins: evolutionarily conserved mediators of immune responses. *Annu Rev Immunol*, **16**, 225-260.

Gibcus, J. H. and Dekker, J. (2013). The hierarchy of the 3D genome. *Mol Cell*, **49**, (5), 773-782.

Gnirke, A., Melnikov, A., Maguire, J., Rogov, P., LeProust, E. M., Brockman, W. *et al.* (2009). Solution hybrid selection with ultra-long oligonucleotides for massively parallel targeted sequencing. *Nat Biotechnol*, **27**, (2), 182-189.

Gottlieb, A. B., Krueger, J. G., Sandberg, L. M., Gothberg, M., and Skolnick, B. E. (2015). First-In-Human, Phase 1, Randomized, Dose-Escalation Trial with Recombinant Anti-IL-20 Monoclonal Antibody in Patients with Psoriasis. *PLoS One*, **10**, (8), e0134703.

Grada, A. and Weinbrecht, K. (2013). Next-generation sequencing: methodology and application. *J Invest Dermatol*, **133**, (8), e11.

Graham, R. R., Cotsapas, C., Davies, L., Hackett, R., Lessard, C. J., Leon, J. M. *et al.* (2008). Genetic variants near TNFAIP3 on 6q23 are associated with systemic lupus erythematosus. *Nat Genet*, **40**, (9), 1059-1061.

Gregersen, P. K. (2005). Pathways to gene identification in rheumatoid arthritis: PTPN22 and beyond. *Immunol Rev*, **204**, 74-86.

Gregersen, P. K., Silver, J., and Winchester, R. J. (1987). The shared epitope hypothesis. An approach to understanding the molecular genetics of susceptibility to rheumatoid arthritis. *Arthritis Rheum*, **30**, (11), 1205-1213.

Griffiths-Jones, S., Grocock, R. J., van, D. S., Bateman, A., and Enright, A. J. (2006). miRBase: microRNA sequences, targets and gene nomenclature. *Nucleic Acids Res*, **34**, (Database issue), D140-D144.

Grundberg, E., Small, K. S., Hedman, A. K., Nica, A. C., Buil, A., Keildson, S. *et al.* (2012). Mapping cis- and trans-regulatory effects across multiple tissues in twins. *Nat Genet*, **44**, (10), 1084-1089.

GTEx Project. (2013). The Genotype-Tissue Expression (GTEx) project. *Nat Genet*, **45**, (6), 580-585.

Guo, Y., Xu, Q., Canzio, D., Shou, J., Li, J., Gorkin, D. U. *et al.* (2015). CRISPR Inversion of CTCF Sites Alters Genome Topology and Enhancer/Promoter Function. *Cell*, **162**, (4), 900-910.

Hagege, H., Klous, P., Braem, C., Splinter, E., Dekker, J., Cathala, G. *et al.* (2007). Quantitative analysis of chromosome conformation capture assays (3C-qPCR). *Nat Protoc*, **2**, (7), 1722-1733.

Haiman, C. A., Patterson, N., Freedman, M. L., Myers, S. R., Pike, M. C., Waliszewska, A. *et al.* (2007). Multiple regions within 8q24 independently affect risk for prostate cancer. *Nat Genet*, **39**, (5), 638-644.

HapMap (2003). The International HapMap Project. *Nature*, **426**, (6968), 789-796.

Hawkins, R. D., Larjo, A., Tripathi, S. K., Wagner, U., Luu, Y., Lonnberg, T. *et al.* (2013). Global chromatin state analysis reveals lineage-specific enhancers during the initiation of human T helper 1 and T helper 2 cell polarization. *Immunity*, **38**, (6), 1271-1284.

Hawtree, S., Muthana, M., Wilkinson, J. M., Akil, M., and Wilson, A. G. (2015). Histone deacetylase 1 regulates tissue destruction in rheumatoid arthritis. *Hum Mol Genet*, **24**, (19), 5367-5377.

Heidari, N., Phanstiel, D. H., He, C., Grubert, F., Jahanbani, F., Kasowski, M. *et al.* (2014). Genome-wide map of regulatory interactions in the human genome. *Genome Res*, **24**, (12), 1905-1917.

Heintzman, N. D., Stuart, R. K., Hon, G., Fu, Y., Ching, C. W., Hawkins, R. D. *et al.* (2007). Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. *Nat Genet*, **39**, (3), 311-318.

Hellman, L. M. and Fried, M. G. (2007). Electrophoretic mobility shift assay (EMSA) for detecting protein-nucleic acid interactions. *Nat Protoc*, **2**, (8), 1849-1861.

Hinks, A., Cobb, J., Marion, M. C., Prahalad, S., Sudman, M., Bowes, J. *et al.* (2013). Dense genotyping of immune-related disease regions identifies 14 new susceptibility loci for juvenile idiopathic arthritis. *Nat Genet*, **45**, (6), 664-669.

Hnisz, D., Abraham, B. J., Lee, T. I., Lau, A., Saint-Andre, V., Sigova, A. A. *et al.* (2013). Super-enhancers in the control of cell identity and disease. *Cell*, **155**, (4), 934-947.

Hnisz, D., Weintraub, A. S., Day, D. S., Valton, A. L., Bak, R. O., Li, C. H. *et al.* (2016). Activation of proto-oncogenes by disruption of chromosome neighborhoods. *Science*, **351**, (6280), 1454-1458.

Hou, W. S., Li, Z., Gordon, R. E., Chan, K., Klein, M. J., Levy, R. *et al.* (2001). Cathepsin k is a critical protease in synovial fibroblast-mediated collagen degradation. *Am J Pathol*, **159**, (6), 2167-2177.

Hsu, Y. H. and Chang, M. S. (2015). IL-20 in rheumatoid arthritis. *Drug Discov Today*.

Hsu, Y. H., Chiu, Y. S., Chen, W. Y., Huang, K. Y., Jou, I. M., Wu, P. T. *et al.* (2016). Anti-IL-20 monoclonal antibody promotes bone fracture healing through regulating IL-20-mediated osteoblastogenesis. *Sci Rep*, **6**, 24339.

Hu, F., Li, Y., Zheng, L., Shi, L., Liu, H., Zhang, X. *et al.* (2014). Toll-like receptors expressed by synovial fibroblasts perpetuate Th1 and th17 cell responses in rheumatoid arthritis. *PLoS One*, **9**, (6), e100266.

Hu, N., Qiu, X., Luo, Y., Yuan, J., Li, Y., Lei, W. *et al.* (2008a). Abnormal histone modification patterns in lupus CD4+ T cells. *J Rheumatol*, **35**, (5), 804-810.

Hu, X., Chakravarty, S. D., and Ivashkiv, L. B. (2008b). Regulation of interferon and Toll-like receptor signaling during macrophage activation by opposing feedforward and feedback inhibition mechanisms. *Immunol Rev*, **226**, 41-56.

Huber, L. C., Distler, O., Tarner, I., Gay, R. E., Gay, S., and Pap, T. (2006). Synovial fibroblasts: key players in rheumatoid arthritis. *Rheumatology (Oxford)*, **45**, (6), 669-675.

Hubner, M. R., Eckersley-Maslin, M. A., and Spector, D. L. (2013). Chromatin organization and transcriptional regulation. *Curr Opin Genet Dev*, **23**, (2), 89-95.

Hughes, J. R., Roberts, N., McGowan, S., Hay, D., Giannoulatou, E., Lynch, M. *et al.* (2014). Analysis of hundreds of cis-regulatory landscapes at high resolution in a single, high-throughput experiment. *Nat Genet*, **46**, (2), 205-212.

Jackson, V. (1999). Formaldehyde cross-linking for studying nucleosomal dynamics. *Methods*, **17**, (2), 125-139.

Jager, R., Migliorini, G., Henrion, M., Kandaswamy, R., Speedy, H. E., Heindl, A. *et al.* (2015). Capture Hi-C identifies the chromatin interactome of colorectal cancer risk loci. *Nat Commun*, **6**, 6178.

Ji, J. D., Lee, W. J., Kong, K. A., Woo, J. H., Choi, S. J., Lee, Y. H. *et al.* (2010). Association of STAT4 polymorphism with rheumatoid arthritis and systemic lupus erythematosus: a meta-analysis. *Mol Biol Rep*, **37**, (1), 141-147.

Jin, F., Li, Y., Dixon, J. R., Selvaraj, S., Ye, Z., Lee, A. Y. *et al.* (2013). A high-resolution map of the three-dimensional chromatin interactome in human cells. *Nature*, **503**, (7475), 290-294.

Kalhor, R., Tjong, H., Jayathilaka, N., Alber, F., and Chen, L. (2012). Genome architectures revealed by tethered chromosome conformation capture and population-based modeling. *Nat Biotechnol*, **30**, (1), 90-98.

Kallberg, H., Ding, B., Padyukov, L., Bengtsson, C., Ronnelid, J., Klareskog, L. *et al.* (2011). Smoking is a major preventable risk factor for rheumatoid arthritis: estimations of risks after various exposures to cigarette smoke. *Ann Rheum Dis*, **70**, (3), 508-511.

Karlson, E. W. and Costenbader, K. H. (2010). Epidemiology: Interpreting studies of interactions between RA risk factors. *Nat Rev Rheumatol*, **6**, (2), 72-73.

Karouzakis, E., Gay, R. E., Gay, S., and Neidhart, M. (2009). Epigenetic control in rheumatoid arthritis synovial fibroblasts. *Nat Rev Rheumatol*, **5**, (5), 266-272.

Kilpinen, H., Waszak, S. M., Gschwind, A. R., Raghav, S. K., Witwicki, R. M., Orioli, A. *et al.* (2013). Coordinated effects of sequence variation on DNA binding, chromatin structure, and transcription. *Science*, **342**, (6159), 744-747.

Klareskog, L., Catrina, A. I., and Paget, S. (2009). Rheumatoid arthritis. *Lancet*, **373**, (9664), 659-672.

Klareskog, L., Malmstrom, V., Lundberg, K., Padyukov, L., and Alfredsson, L. (2011). Smoking, citrullination and genetic variability in the immunopathogenesis of rheumatoid arthritis. *Semin Immunol*, **23**, (2), 92-98.

Klein, K. and Gay, S. (2015). Epigenetics in rheumatoid arthritis. *Curr Opin Rheumatol*, **27**, (1), 76-82.

Klein, K., Ospelt, C., and Gay, S. (2012). Epigenetic contributions in the development of rheumatoid arthritis. *Arthritis Res Ther*, **14**, (6), 227.

Klockars, M., Koskela, R. S., Jarvinen, E., Kolari, P. J., and Rossi, A. (1987). Silica exposure and rheumatoid arthritis: a follow up study of granite workers 1940-81. *Br Med J (Clin Res Ed)*, **294**, (6578), 997-1000.

Knight, J. C., Keating, B. J., Rockett, K. A., and Kwiatkowski, D. P. (2003). In vivo characterization of regulatory polymorphisms by allele-specific quantification of RNA polymerase loading. *Nat Genet*, **33**, (4), 469-475.

Kolovos, P., van de Werken, H. J., Kepper, N., Zuin, J., Brouwer, R. W., Kockx, C. E. *et al.* (2014). Targeted Chromatin Capture (T2C): a novel high resolution high throughput method to detect genomic interactions and regulatory elements. *Epigenetics Chromatin*, **7**, 10.

Kouzarides, T. (2002). Histone methylation in transcriptional control. *Curr Opin Genet Dev*, **12**, (2), 198-209.

Kouzarides, T. (2007). Chromatin modifications and their function. *Cell*, **128**, (4), 693-705.

Kumar, V., Wijmenga, C., and Xavier, R. J. (2014). Genetics of immune-mediated disorders: from genome-wide association to molecular mechanism. *Curr Opin Immunol*, **31**, 51-57.

Kurko, J., Besenyei, T., Laki, J., Glant, T. T., Mikecz, K., and Szekanecz, Z. (2013). Genetics of rheumatoid arthritis - a comprehensive review. *Clin Rev Allergy Immunol*, **45**, (2), 170-179.

Lafyatis, R., Remmers, E. F., Roberts, A. B., Yocum, D. E., Sporn, M. B., and Wilder, R. L. (1989). Anchorage-independent growth of synoviocytes from arthritic and normal joints. Stimulation by exogenous platelet-derived growth factor and inhibition by transforming growth factor-beta and retinoids. *J Clin Invest*, **83**, (4), 1267-1276.

Lam, M. T., Li, W., Rosenfeld, M. G., and Glass, C. K. (2014). Enhancer RNAs and regulated transcriptional programs. *Trends Biochem Sci*, **39**, (4), 170-182.

Lamas, J. R., Rodriguez-Rodriguez, L., Varade, J., Lopez-Romero, P., Tornero-Esteban, P., Abasolo, L. *et al.* (2010). Influence of IL6R rs8192284 polymorphism status in disease activity in rheumatoid arthritis. *J Rheumatol*, **37**, (8), 1579-1581.

Lanctot, C., Cheutin, T., Cremer, M., Cavalli, G., and Cremer, T. (2007). Dynamic genome architecture in the nuclear space: regulation of gene expression in three dimensions. *Nat Rev Genet*, **8**, (2), 104-115.

Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J. *et al.* (2001). Initial sequencing and analysis of the human genome. *Nature*, **409**, (6822), 860-921.

Langmead, B., Trapnell, C., Pop, M., and Salzberg, S. L. (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol*, **10**, (3), R25.

Lee, M. N., Ye, C., Villani, A. C., Raj, T., Li, W., Eisenhaure, T. M. *et al.* (2014). Common genetic variants modulate pathogen-sensing responses in human dendritic cells. *Science*, **343**, (6175), 1246980.

Lei, W., Luo, Y., Lei, W., Luo, Y., Yan, K., Zhao, S. *et al.* (2009). Abnormal DNA methylation in CD4+ T cells from patients with systemic lupus erythematosus, systemic sclerosis, and dermatomyositis. *Scand J Rheumatol*, **38**, (5), 369-374.

Levine, M., Cattoglio, C., and Tjian, R. (2014). Looping back to leap forward: transcription enters a new era. *Cell*, **157**, (1), 13-25.

Li, J., Humphreys, K., Heikkinen, T., Aittomaki, K., Blomqvist, C., Pharoah, P. D. *et al.* (2011). A combined analysis of genome-wide association studies in breast cancer. *Breast Cancer Res Treat*, **126**, (3), 717-727.

Li, Y., Rivera, C. M., Ishii, H., Jin, F., Selvaraj, S., Lee, A. Y. *et al.* (2014). CRISPR reveals a distal super-enhancer required for Sox2 expression in mouse embryonic stem cells. *PLoS One*, **9**, (12), e114485.

Liao, J., Liang, G., Xie, S., Zhao, H., Zuo, X., Li, F. *et al.* (2012). CD40L demethylation in CD4(+) T cells from women with rheumatoid arthritis. *Clin Immunol*, **145**, (1), 13-18.

Lieberman-Aiden, E., van Berkum, N. L., Williams, L., Imakaev, M., Ragoczy, T., Telling, A. *et al.* (2009). Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science*, **326**, (5950), 289-293.

Liu, Y., Aryee, M. J., Padyukov, L., Fallin, M. D., Hesselberg, E., Runarsson, A. *et al.* (2013). Epigenome-wide association data implicate DNA methylation as an intermediary of genetic risk in rheumatoid arthritis. *Nat Biotechnol*, **31**, (2), 142-147.

Liu, Z., Merkurjev, D., Yang, F., Li, W., Oh, S., Friedman, M. J. *et al.* (2014). Enhancer activation requires trans-recruitment of a mega transcription factor complex. *Cell*, **159**, (2), 358-373.

Lower, K. M., Hughes, J. R., De, G. M., Henderson, S., Viprakasit, V., Fisher, C. *et al.* (2009). Adventitious changes in long-range gene expression caused by polymorphic structural variation and promoter competition. *Proc Natl Acad Sci U S A*, **106**, (51), 21771-21776.

Loyola-Rodriguez, J. P., Martinez-Martinez, R. E., Abud-Mendoza, C., Patino-Marin, N., and Seymour, G. J. (2010). Rheumatoid arthritis and the role of oral bacteria. *J Oral Microbiol*, **2**.

Lu, Q. (2013). The critical importance of epigenetics in autoimmunity. *J Autoimmun*, **41**, 1-5.

Luckey, D., Medina, K., and Taneja, V. (2012). B cells as effectors and regulators of sex-biased arthritis. *Autoimmunity*, **45**, (5), 364-376.

Lupianez, D. G., Kraft, K., Heinrich, V., Krawitz, P., Brancati, F., Klopocki, E. *et al.* (2015). Disruptions of topological chromatin domains cause pathogenic rewiring of gene-enhancer interactions. *Cell*, **161**, (5), 1012-1025.

Ma, W., Ay, F., Lee, C., Gulsoy, G., Deng, X., Cook, S. *et al.* (2015). Fine-scale chromatin interaction maps reveal the cis-regulatory landscape of human lincRNA genes. *Nat Methods*, **12**, (1), 71-78.

Macgregor, A. J., Snieder, H., Rigby, A. S., Koskenvuo, M., Kaprio, J., Aho, K. *et al.* (2000). Characterizing the quantitative genetic contribution to rheumatoid arthritis using data from twins. *Arthritis Rheum*, **43**, (1), 30-37.

Majewski, J. and Pastinen, T. (2011). The study of eQTL variations by RNA-seq: from SNPs to phenotypes. *Trends Genet*, **27**, (2), 72-79.

Mardis, E. R. (2013). Next-generation sequencing platforms. *Annu Rev Anal Chem (Palo Alto Calif)*, **6**, 287-303.

Martens, J. H. and Stunnenberg, H. G. (2013). BLUEPRINT: mapping human blood cell epigenomes. *Haematologica*, **98**, (10), 1487-1489.

Martin, P., McGovern, A., Orozco, G., Duffus, K., Yarwood, A., Schoenfelder, S. *et al.* (2015). Capture Hi-C reveals novel candidate genes and complex long-range interactions with related autoimmune risk loci. *Nat Commun*, **6**, 10069.

Masi, A. T., Aldag, J. C., and Chatterton, R. T. (2006). Sex hormones and risks of rheumatoid arthritis and developmental or environmental influences. *Ann N Y Acad Sci*, **1069**, 223-235.

Maston, G. A., Evans, S. K., and Green, M. R. (2006). Transcriptional regulatory elements in the human genome. *Annu Rev Genomics Hum Genet*, **7**, 29-59.

Maurano, M. T., Haugen, E., Sandstrom, R., Vierstra, J., Shafer, A., Kaul, R. *et al.* (2015). Large-scale identification of sequence variants influencing human transcription factor occupancy in vivo. *Nat Genet*, **47**, (12), 1393-1401.

Maurano, M. T., Humbert, R., Rynes, E., Thurman, R. E., Haugen, E., Wang, H. *et al.* (2012). Systematic localization of common disease-associated variation in regulatory DNA. *Science*, **337**, (6099), 1190-1195.

McAllister, K., Yarwood, A., Bowes, J., Orozco, G., Viatte, S., Diogo, D. *et al.* (2013). Identification of BACH2 and RAD51B as rheumatoid arthritis susceptibility loci in a meta-analysis of genome-wide data. *Arthritis Rheum*, **65**, (12), 3058-3062.

McGarry, T., Veale, D. J., Gao, W., Orr, C., Fearon, U., and Connolly, M. (2015). Toll-like receptor 2 (TLR2) induces migration and invasive mechanisms in rheumatoid arthritis. *Arthritis Res Ther*, **17**, 153.

McInnes, I. B. and Schett, G. (2007). Cytokines in the pathogenesis of rheumatoid arthritis. *Nat Rev Immunol*, **7**, (6), 429-442.

McInnes, I. B. and Schett, G. (2011). The pathogenesis of rheumatoid arthritis. *N Engl J Med*, **365**, (23), 2205-2219.

Mercier, E., Droit, A., Li, L., Robertson, G., Zhang, X., and Gottardo, R. (2011). An integrated pipeline for the genome-wide analysis of transcription factor binding sites from ChIP-Seq. *PLoS One*, **6**, (2), e16432.

Metivier, R., Penot, G., Hubner, M. R., Reid, G., Brand, H., Kos, M. *et al.* (2003). Estrogen receptor-alpha directs ordered, cyclical, and combinatorial recruitment of cofactors on a natural target promoter. *Cell*, **115**, (6), 751-763.

Meyer, K. B., Maia, A. T., O'Reilly, M., Ghoussaini, M., Prathalingam, R., Porter-Gill, P. *et al.* (2011). A functional variant at a prostate cancer predisposition locus at 8q24 is associated with PVT1 expression. *PLoS Genet*, **7**, (7), e1002165.

Meyer, K. B., O'Reilly, M., Michailidou, K., Carlebur, S., Edwards, S. L., French, J. D. *et al.* (2013). Fine-scale mapping of the FGFR2 breast cancer risk locus: putative functional variants differentially bind FOXA1 and E2F1. *Am J Hum Genet*, **93**, (6), 1046-1060.

Miele, A., Gheldof, N., Tabuchi, T. M., Dostie, J., and Dekker, J. (2006). Mapping chromatin interactions by chromosome conformation capture. *Curr Protoc Mol Biol*, **Chapter 21**, Unit.

Mifsud, B., Tavares-Cadete, F., Young, A. N., Sugar, R., Schoenfelder, S., Ferreira, L. *et al.* (2015). Mapping long-range promoter contacts in human cells with high-resolution capture Hi-C. *Nat Genet*, **47**, (6), 598-606.

Monaco, C., Andreakos, E., Kiriakidis, S., Feldmann, M., and Paleolog, E. (2004a). T-cell-mediated signalling in immune, inflammatory and angiogenic processes: the cascade of events leading to inflammatory diseases. *Curr Drug Targets Inflamm Allergy*, **3**, (1), 35-42.

Monaco, C., Andreakos, E., Kiriakidis, S., Mauri, C., Bicknell, C., Foxwell, B. *et al.* (2004b). Canonical pathway of nuclear factor kappa B activation selectively regulates proinflammatory and prothrombotic responses in human atherosclerosis. *Proc Natl Acad Sci U S A*, **101**, (15), 5634-5639.

Montgomery, S. B., Sammeth, M., Gutierrez-Arcelus, M., Lach, R. P., Ingle, C., Nisbett, J. *et al.* (2010). Transcriptome genetics using second generation sequencing in a Caucasian population. *Nature*, **464**, (7289), 773-777.

Morgan, A. W., Robinson, J. I., Conaghan, P. G., Martin, S. G., Hensor, E. M., Morgan, M. D. *et al.* (2010). Evaluation of the rheumatoid arthritis susceptibility loci HLA-DRB1, PTPN22, OLIG3/TNFAIP3, STAT4 and TRAF1/C5 in an inception cohort. *Arthritis Res Ther*, **12**, (2), R57.

Mousavi, K., Zare, H., Koulnis, M., and Sartorelli, V. (2014). The emerging roles of eRNAs in transcriptional regulatory networks. *RNA Biol*, **11**, (2), 106-110.

Moynagh, P. N. (2005). The NF-kappaB pathway. *J Cell Sci*, **118**, (Pt 20), 4589-4592.

Muller-Ladner, U., Kriegsmann, J., Gay, R. E., and Gay, S. (1995). Oncogenes in rheumatoid arthritis. *Rheum Dis Clin North Am*, **21**, (3), 675-690.

Musunuru, K., Strong, A., Frank-Kamenetsky, M., Lee, N. E., Ahfeldt, T., Sachs, K. V. *et al.* (2010). From noncoding variant to phenotype via SORT1 at the 1p13 cholesterol locus. *Nature*, **466**, (7307), 714-719.

Nagano, T., Lubling, Y., Stevens, T. J., Schoenfelder, S., Yaffe, E., Dean, W. *et al.* (2013). Single-cell Hi-C reveals cell-to-cell variability in chromosome structure. *Nature*, **502**, (7469), 59-64.

Nagano, T., Varnai, C., Schoenfelder, S., Javierre, B. M., Wingett, S. W., and Fraser, P. (2015). Comparison of Hi-C results using in-solution versus in-nucleus ligation. *Genome Biol*, **16**, (1), 175.

Nakano, K., Boyle, D. L., and Firestein, G. S. (2013). Regulation of DNA methylation in rheumatoid arthritis synoviocytes. *J Immunol*, **190**, (3), 1297-1303.

Naumova, N., Imakaev, M., Fudenberg, G., Zhan, Y., Lajoie, B. R., Mirny, L. A. *et al.* (2013). Organization of the mitotic chromosome. *Science*, **342**, (6161), 948-953.

Naumova, N., Smith, E. M., Zhan, Y., and Dekker, J. (2012). Analysis of long-range chromatin interactions using Chromosome Conformation Capture. *Methods*, **58**, (3), 192-203.

Nelson, M. R., Tipney, H., Painter, J. L., Shen, J., Nicoletti, P., Shen, Y. *et al.* (2015). The support of human genetic evidence for approved drug indications. *Nat Genet*, **47**, (8), 856-860.

Nica, A. C., Montgomery, S. B., Dimas, A. S., Stranger, B. E., Beazley, C., Barroso, I. *et al.* (2010). Candidate causal regulatory effects by integration of expression QTLs with complex trait genetic associations. *PLoS Genet*, **6**, (4), e1000895.

Nica, A. C., Parts, L., Glass, D., Nisbet, J., Barrett, A., Sekowska, M. *et al.* (2011). The architecture of gene regulatory variation across multiple human tissues: the MuTHER study. *PLoS Genet*, **7**, (2), e1002003.

Nicolae, D. L., Gamazon, E., Zhang, W., Duan, S., Dolan, M. E., and Cox, N. J. (2010). Trait-associated SNPs are more likely to be eQTLs: annotation to enhance discovery from GWAS. *PLoS Genet*, **6**, (4), e1000888.

Nile, C. J., Read, R. C., Akil, M., Duff, G. W., and Wilson, A. G. (2008). Methylation status of a single CpG site in the IL6 promoter is related to IL6 messenger RNA levels and rheumatoid arthritis. *Arthritis Rheum*, **58**, (9), 2686-2693.

O'Neill, L. P. and Turner, B. M. (2003). Immunoprecipitation of native chromatin: NChIP. *Methods*, **31**, (1), 76-82.

Okada, Y., Terao, C., Ikari, K., Kochi, Y., Ohmura, K., Suzuki, A. *et al.* (2012). Meta-analysis identifies nine new loci associated with rheumatoid arthritis in the Japanese population. *Nat Genet*, **44**, (5), 511-516.

Okada, Y., Wu, D., Trynka, G., Raj, T., Terao, C., Ikari, K. *et al.* (2014). Genetics of rheumatoid arthritis contributes to biology and drug discovery. *Nature*, **506**, (7488), 376-381.

Onengut-Gumuscu, S., Chen, W. M., Burren, O., Cooper, N. J., Quinlan, A. R., Mychaleckyj, J. C. *et al.* (2015). Fine mapping of type 1 diabetes susceptibility loci and evidence for colocalization of causal variants with lymphoid gene enhancers. *Nat Genet*, **47**, (4), 381-386.

Ong, C. T. and Corces, V. G. (2014). CTCF: an architectural protein bridging genome topology and function. *Nat Rev Genet*, **15**, (4), 234-246.

Orozco, G., Alizadeh, B. Z., Delgado-Vega, A. M., Gonzalez-Gay, M. A., Balsa, A., Pascual-Salcedo, D. *et al.* (2008). Association of STAT4 with rheumatoid arthritis: a replication study in three European populations. *Arthritis Rheum*, **58**, (7), 1974-1980.

Orozco, G., Hinks, A., Eyre, S., Ke, X., Gibbons, L. J., Bowes, J. *et al.* (2009). Combined effects of three independent SNPs greatly increase the risk estimate for RA at 6q23. *Hum Mol Genet*, **18**, (14), 2693-2699.

Orozco, G., Viatte, S., Bowes, J., Martin, P., Wilson, A. G., Morgan, A. W. *et al.* (2014). Novel rheumatoid arthritis susceptibility locus at 22q12 identified in an extended UK genome-wide association study. *Arthritis Rheumatol*, **66**, (1), 24-30.

Osborne, C. S., Chakalova, L., Brown, K. E., Carter, D., Horton, A., Debrand, E. *et al.* (2004). Active genes dynamically colocalize to shared sites of ongoing transcription. *Nat Genet*, **36**, (10), 1065-1071.

Palstra, R. J., Tolhuis, B., Splinter, E., Nijmeijer, R., Grosveld, F., and de, L. W. (2003). The beta-globin nuclear compartment in development and erythroid differentiation. *Nat Genet*, **35**, (2), 190-194.

Pan, W., Zhu, S., Yuan, M., Cui, H., Wang, L., Luo, X. *et al.* (2010). MicroRNA-21 and microRNA-148a contribute to DNA hypomethylation in lupus CD4+ T cells by directly and indirectly targeting DNA methyltransferase 1. *J Immunol*, **184**, (12), 6773-6781.

Park, S. H., Kim, S. K., Choe, J. Y., Moon, Y., An, S., Park, M. J. *et al.* (2013). Hypermethylation of EBF3 and IRX1 genes in synovial fibroblasts of patients with rheumatoid arthritis. *Mol Cells*, **35**, (4), 298-304.

Parker, S. C., Stitzel, M. L., Taylor, D. L., Orozco, J. M., Erdos, M. R., Akiyama, J. A. *et al.* (2013). Chromatin stretch enhancer states drive cell-specific gene regulation and harbor human disease risk variants. *Proc Natl Acad Sci U S A*, **110**, (44), 17921-17926.

Pennacchio, L. A., Bickmore, W., Dean, A., Nobrega, M. A., and Bejerano, G. (2013). Enhancers: five essential questions. *Nat Rev Genet*, **14**, (4), 288-295.

Pestka, S., Krause, C. D., Sarkar, D., Walter, M. R., Shi, Y., and Fisher, P. B. (2004). Interleukin-10 and related cytokines and receptors. *Annu Rev Immunol*, **22**, 929-979.

Pham, H., Kearns, N. A., and Maehr, R. (2016). Transcriptional Regulation with CRISPR/Cas9 Effectors in Mammalian Cells. *Methods Mol Biol*, **1358**, 43-57.

Phillips-Cremins, J. E., Sauria, M. E., Sanyal, A., Gerasimova, T. I., Lajoie, B. R., Bell, J. S. *et al.* (2013). Architectural protein subclasses shape 3D organization of genomes during lineage commitment. *Cell*, **153**, (6), 1281-1295.

Pickrell, J. K., Marioni, J. C., Pai, A. A., Degner, J. F., Engelhardt, B. E., Nkadori, E. *et al.* (2010). Understanding mechanisms underlying human gene expression variation with RNA sequencing. *Nature*, **464**, (7289), 768-772.

Pittman, A. M., Naranjo, S., Jalava, S. E., Twiss, P., Ma, Y., Olver, B. *et al.* (2010). Allelic variation at the 8q23.3 colorectal cancer risk locus functions as a cis-acting regulator of EIF3H. *PLoS Genet*, **6**, (9), e1001126.

Plenge, R. M., Cotsapas, C., Davies, L., Price, A. L., de Bakker, P. I., Maller, J. *et al.* (2007). Two independent alleles at 6q23 associated with risk of rheumatoid arthritis. *Nat Genet*, **39**, (12), 1477-1482.

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

Pott, S. and Lieb, J. D. (2015). What are super-enhancers? *Nat Genet*, **47**, (1), 8-12.

Quail, M. A., Smith, M., Coupland, P., Otto, T. D., Harris, S. R., Connor, T. R. *et al.* (2012). A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers. *BMC Genomics*, **13**, 341.

Quinlan, A. R. and Hall, I. M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, **26**, (6), 841-842.

Rao, S. S., Huntley, M. H., Durand, N. C., Stamenova, E. K., Bochkov, I. D., Robinson, J. T. *et al.* (2014). A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell*, **159**, (7), 1665-1680.

Raychaudhuri, S., Sandor, C., Stahl, E. A., Freudenberg, J., Lee, H. S., Jia, X. *et al.* (2012). Five amino acids in three HLA proteins explain most of the association between MHC and seropositive rheumatoid arthritis. *Nat Genet*, **44**, (3), 291-296.

Reich, D. E. and Lander, E. S. (2001). On the allelic spectrum of human disease. *Trends Genet*, **17**, (9), 502-510.

Remmers, E. F., Plenge, R. M., Lee, A. T., Graham, R. R., Hom, G., Behrens, T. W. *et al.* (2007). STAT4 and the risk of rheumatoid arthritis and systemic lupus erythematosus. *N Engl J Med*, **357**, (10), 977-986.

Rhee, H. S. and Pugh, B. F. (2012). ChIP-exo method for identifying genomic location of DNA-binding proteins with near-single-nucleotide accuracy. *Curr Protoc Mol Biol*, **Chapter 21**, Unit.

Ricano-Ponce, I. and Wijmenga, C. (2013). Mapping of immune-mediated disease genes. *Annu Rev Genomics Hum Genet*, **14**, 325-353.

Ricano-Ponce, I., Zhernakova, D. V., Deelen, P., Luo, O., Li, X., Isaacs, A. *et al.* (2016). Refined mapping of autoimmune disease associated genetic variants with gene expression suggests an important role for non-coding RNAs. *J Autoimmun*, **68**, 62-74.

Richardson, B., Scheinbart, L., Strahler, J., Gross, L., Hanash, S., and Johnson, M. (1990). Evidence for impaired T cell DNA methylation in systemic lupus erythematosus and rheumatoid arthritis. *Arthritis Rheum*, **33**, (11), 1665-1673.

Rinn, J. L. and Chang, H. Y. (2012). Genome regulation by long noncoding RNAs. *Annu Rev Biochem*, **81**, 145-166.

Rivera, C. M. and Ren, B. (2013). Mapping human epigenomes. *Cell*, **155**, (1), 39-55.

Robyr, D., Friedli, M., Gehrig, C., Arcangeli, M., Marin, M., Guipponi, M. *et al.* (2011). Chromosome conformation capture uncovers potential genome-wide interactions between human conserved non-coding sequences. *PLoS One*, **6**, (3), e17634.

Rodriguez, A. and Bjerling, P. (2013). The links between chromatin spatial organization and biological function. *Biochem Soc Trans*, **41**, (6), 1634-1639.

Rosenfeld, M. G., Lunyak, V. V., and Glass, C. K. (2006). Sensors and signals: a coactivator/corepressor/epigenetic code for integrating signal-dependent programs of transcriptional response. *Genes Dev*, **20**, (11), 1405-1428.

Rutz, S., Wang, X., and Ouyang, W. (2014). The IL-20 subfamily of cytokines--from host defence to tissue homeostasis. *Nat Rev Immunol*, **14**, (12), 783-795.

Sabeh, F., Fox, D., and Weiss, S. J. (2010). Membrane-type I matrix metalloproteinase-dependent regulation of rheumatoid arthritis synoviocyte function. *J Immunol*, **184**, (11), 6396-6406.

Sachidanandam, R., Weissman, D., Schmidt, S. C., Kakol, J. M., Stein, L. D., Marth, G. *et al.* (2001). A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. *Nature*, **409**, (6822), 928-933.

Sahlen, P., Abdullayev, I., Ramskold, D., Matskova, L., Rilakovic, N., Lotstedt, B. *et al.* (2015). Genome-wide mapping of promoter-anchored interactions with close to single-enhancer resolution. *Genome Biol*, **16**, 156.

Sakurai, N., Kuroiwa, T., Ikeuchi, H., Hiramatsu, N., Maeshima, A., Kaneko, Y. *et al.* (2008). Expression of IL-19 and its receptors in RA: potential role for synovial hyperplasia formation. *Rheumatology (Oxford)*, **47**, (6), 815-820.

Sanger, F., Nicklen, S., and Coulson, A. R. (1977). DNA sequencing with chain-terminating inhibitors. *Proc Natl Acad Sci U S A*, **74**, (12), 5463-5467.

Sanyal, A., Lajoie, B. R., Jain, G., and Dekker, J. (2012). The long-range interaction landscape of gene promoters. *Nature*, **489**, (7414), 109-113.

Scacheri, P. C., Crawford, G. E., and Davis, S. (2006). Statistics for ChIP-chip and DNase hypersensitivity experiments on NimbleGen arrays. *Methods Enzymol*, **411**, 270-282.

Schaub, M. A., Boyle, A. P., Kundaje, A., Batzoglou, S., and Snyder, M. (2012). Linking disease associations with regulatory information in the human genome. *Genome Res*, **22**, (9), 1748-1759.

Scher, J. U. and Abramson, S. B. (2011). The microbiome and rheumatoid arthritis. *Nat Rev Rheumatol*, **7**, (10), 569-578.

Schett, G. and Gravallese, E. (2012). Bone erosion in rheumatoid arthritis: mechanisms, diagnosis and treatment. *Nat Rev Rheumatol*, **8**, (11), 656-664.

Schneider, U., Schwenk, H. U., and Bornkamm, G. (1977). 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**, (5), 621-626.

Schodel, J., Bardella, C., Sciesielski, L. K., Brown, J. M., Pugh, C. W., Buckle, V. *et al.* (2012). Common genetic variants at the 11q13.3 renal cancer susceptibility locus influence binding of HIF to an enhancer of cyclin D1 expression. *Nat Genet*, **44**, (4), 420-422.

Schoenfelder, S., Clay, I., and Fraser, P. (2010a). The transcriptional interactome: gene expression in 3D. *Curr Opin Genet Dev*, **20**, (2), 127-133.

Schoenfelder, S., Furlan-Magaril, M., Mifsud, B., Tavares-Cadete, F., Sugar, R., Javierre, B. M. *et al.* (2015). The pluripotent regulatory circuitry connecting promoters to their long-range interacting elements. *Genome Res*, **25**, (4), 582-597.

Schoenfelder, S., Sexton, T., Chakalova, L., Cope, N. F., Horton, A., Andrews, S. *et al.* (2010b). Preferential associations between co-regulated genes reveal a transcriptional interactome in erythroid cells. *Nat Genet*, **42**, (1), 53-61.

Senolt, L., Leszczynski, P., Dokoupilova, E., Gothberg, M., Valencia, X., Hansen, B. B. *et al.* (2015). Efficacy and Safety of Anti-Interleukin-20 Monoclonal Antibody in Patients With Rheumatoid Arthritis: A Randomized Phase IIa Trial. *Arthritis Rheumatol*, **67**, (6), 1438-1448.

Sexton, T., Bantignies, F., and Cavalli, G. (2009). Genomic interactions: chromatin loops and gene meeting points in transcriptional regulation. *Semin Cell Dev Biol*, **20**, (7), 849-855.

Sexton, T., Kurukuti, S., Mitchell, J. A., Umlauf, D., Nagano, T., and Fraser, P. (2012a). Sensitive detection of chromatin coassociations using enhanced chromosome conformation capture on chip. *Nat Protoc*, **7**, (7), 1335-1350.

Sexton, T., Yaffe, E., Kenigsberg, E., Bantignies, F., Leblanc, B., Hoichman, M. *et al.* (2012b). Three-dimensional folding and functional organization principles of the Drosophila genome. *Cell*, **148**, (3), 458-472.

Sikes, M. L., Bradshaw, J. M., Ivory, W. T., Lunsford, J. L., McMillan, R. E., and Morrison, C. R. (2009). A streamlined method for rapid and sensitive chromatin immunoprecipitation. *J Immunol Methods*, **344**, (1), 58-63.

Silman, A. J. (1993). Smoking and the risk of rheumatoid arthritis. *J Rheumatol*, **20**, (11), 1815-1816.

Silman, A. J., Macgregor, A. J., Thomson, W., Holligan, S., Carthy, D., Farhan, A. *et al.* (1993). Twin concordance rates for rheumatoid arthritis: results from a nationwide study. *Br J Rheumatol*, **32**, (10), 903-907.

Simonis, M., Klous, P., Splinter, E., Moshkin, Y., Willemsen, R., de, W. E. *et al.* (2006). Nuclear organization of active and inactive chromatin domains uncovered by chromosome conformation capture-on-chip (4C). *Nat Genet*, **38**, (11), 1348-1354.

Simonis, M., Kooren, J., and de, L. W. (2007). An evaluation of 3C-based methods to capture DNA interactions. *Nat Methods*, **4**, (11), 895-901.

Singh, J. A., Furst, D. E., Bharat, A., Curtis, J. R., Kavanaugh, A. F., Kremer, J. M. *et al.* (2012). 2012 update of the 2008 American College of Rheumatology recommendations for the use of disease-modifying antirheumatic drugs and biologic agents in the treatment of rheumatoid arthritis. *Arthritis Care Res (Hoboken)*, **64**, (5), 625-639.

Smallwood, A. and Ren, B. (2013). Genome organization and long-range regulation of gene expression by enhancers. *Curr Opin Cell Biol*, **25**, (3), 387-394.

Smith, M. D. and Tak, P. P. (2002). Rheumatoid arthritis: new insights into the role of synovial inflammation in joint destruction. *Mod Rheumatol*, **12**, (4), 287-293.

Smolen, J. S. and et al (2010). EULAR recommendations for the management of rheumatoid arthritis with synthetic and biological disease-modifying antirheumatic drugs. *Ann Rheum Dis*, **69**, (6), 964-975.

Soldner, F., Stelzer, Y., Shivalila, C. S., Abraham, B. J., Latourelle, J. C., Barrasa, M. I. *et al.* (2016). Parkinson-associated risk variant in distal enhancer of alpha-synuclein modulates target gene expression. *Nature*, **533**, (7601), 95-99.

Sotelo, J., Esposito, D., Duhagon, M. A., Banfield, K., Mehalko, J., Liao, H. *et al.* (2010). Long-range enhancers on 8q24 regulate c-Myc. *Proc Natl Acad Sci U S A*, **107**, (7), 3001-3005.

Spilianakis, C. G., Lalioti, M. D., Town, T., Lee, G. R., and Flavell, R. A. (2005). Interchromosomal associations between alternatively expressed loci. *Nature*, **435**, (7042), 637-645.

Splinter, E., de, W. E., van de Werken, H. J., Klous, P., and de, L. W. (2012). Determining long-range chromatin interactions for selected genomic sites using 4C-seq technology: from fixation to computation. *Methods*, **58**, (3), 221-230.

Splinter, E., Heath, H., Kooren, J., Palstra, R. J., Klous, P., Grosveld, F. *et al.* (2006). CTCF mediates long-range chromatin looping and local histone modification in the beta-globin locus. *Genes Dev*, **20**, (17), 2349-2354.

Stadhouders, R., Kolovos, P., Brouwer, R., Zuin, J., van den Heuvel, A., Kockx, C. *et al.* (2013). Multiplexed chromosome conformation capture sequencing for rapid genome-scale high-resolution detection of long-range chromatin interactions. *Nat Protoc*, **8**, (3), 509-524.

Stahl, E. A., Raychaudhuri, S., Remmers, E. F., Xie, G., Eyre, S., Thomson, B. P. *et al.* (2010). Genome-wide association study meta-analysis identifies seven new rheumatoid arthritis risk loci. *Nat Genet*, **42**, (6), 508-514.

Stanford, S. M. and Bottini, N. (2014). PTPN22: the archetypal non-HLA autoimmunity gene. *Nat Rev Rheumatol*.

Starkebaum, G. (2007). Rheumatoid arthritis and lymphoma: risky business for B cells. *J Rheumatol*, **34**, (2), 243-246.

Stastny, P. and Fink, C. W. (1979). Different HLA-D associations in adult and juvenile rheumatoid arthritis. *J Clin Invest*, **63**, (1), 124-130.

Stolt, P., Kallberg, H., Lundberg, I., Sjogren, B., Klareskog, L., and Alfredsson, L. (2005). Silica exposure is associated with increased risk of developing rheumatoid arthritis: results from the Swedish EIRA study. *Ann Rheum Dis*, **64**, (4), 582-586.

Storch, H., Zimmermann, B., Resch, B., Tykocinski, L. O., Moradi, B., Horn, P. *et al.* (2016). Activated human B cells induce inflammatory fibroblasts with cartilage-destructive properties and become functionally suppressed in return. *Ann Rheum Dis*, **75**, (5), 924-932.

Stranger, B. E., Forrest, M. S., Clark, A. G., Minichiello, M. J., Deutsch, S., Lyle, R. *et al.* (2005). Genome-wide associations of gene expression variation in humans. *PLoS Genet*, **1**, (6), e78.

Stranger, B. E., Montgomery, S. B., Dimas, A. S., Parts, L., Stegle, O., Ingle, C. E. *et al.* (2012). Patterns of cis regulatory variation in diverse human populations. *PLoS Genet*, **8**, (4), e1002639.

Su, A. I., Wiltshire, T., Batalov, S., Lapp, H., Ching, K. A., Block, D. *et al.* (2004). A gene atlas of the mouse and human protein-encoding transcriptomes. *Proc Natl Acad Sci U S A*, **101**, (16), 6062-6067.

Sverdrup, B., Kallberg, H., Bengtsson, C., Lundberg, I., Padyukov, L., Alfredsson, L. *et al.* (2005). Association between occupational exposure to mineral oil and rheumatoid arthritis: results from the Swedish EIRA case-control study. *Arthritis Res Ther*, **7**, (6), R1296-R1303.

Symmons, D. P. (2005). Looking back: rheumatoid arthritis--aetiology, occurrence and mortality. *Rheumatology (Oxford)*, **44 Suppl 4**, iv14-iv17.

Symmons, D. P., Bankhead, C. R., Harrison, B. J., Brennan, P., Barrett, E. M., Scott, D. G. *et al.* (1997). Blood transfusion, smoking, and obesity as risk factors for the development of rheumatoid arthritis: results from a primary care-based incident case-control study in Norfolk, England. *Arthritis Rheum*, **40**, (11), 1955-1961.

Symmons, O., Uslu, V. V., Tsujimura, T., Ruf, S., Nassari, S., Schwarzer, W. *et al.* (2014). Functional and topological characteristics of mammalian regulatory domains. *Genome Res*, **24**, (3), 390-400.

Takami, N., Osawa, K., Miura, Y., Komai, K., Taniguchi, M., Shiraishi, M. *et al.* (2006). Hypermethylated promoter region of DR3, the death receptor 3 gene, in rheumatoid arthritis synovial cells. *Arthritis Rheum*, **54**, (3), 779-787.

Tang, Q., Danila, M. I., Cui, X., Parks, L., Baker, B. J., Reynolds, R. J. *et al.* (2015). 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*, **67**, (5), 1165-1170.

Terao, C., Raychaudhuri, S., and Gregersen, P. K. (2016). Recent Advances in Defining the Genetic Basis of Rheumatoid Arthritis. *Annu Rev Genomics Hum Genet*.

Thomson, W., Barton, A., Ke, X., Eyre, S., Hinks, A., Bowes, J. *et al.* (2007). Rheumatoid arthritis association at 6q23. *Nat Genet*, **39**, (12), 1431-1433.

Tolhuis, B., Palstra, R. J., Splinter, E., Grosveld, F., and de, L. W. (2002). Looping and interaction between hypersensitive sites in the active beta-globin locus. *Mol Cell*, **10**, (6), 1453-1465.

Tomlinson, I. (2012). Colorectal cancer genetics: from candidate genes to GWAS and back again. *Mutagenesis*, **27**, (2), 141-142.

Trenkmann, M., Brock, M., Gay, R. E., Kolling, C., Speich, R., Michel, B. A. *et al.* (2011). Expression and function of EZH2 in synovial fibroblasts: epigenetic repression of the Wnt inhibitor SFRP1 in rheumatoid arthritis. *Ann Rheum Dis*, **70**, (8), 1482-1488.

Trynka, G., Hunt, K. A., Bockett, N. A., Romanos, J., Mistry, V., Szperl, A. *et al.* (2011). Dense genotyping identifies and localizes multiple common and rare variant association signals in celiac disease. *Nat Genet*, **43**, (12), 1193-1201.

Trynka, G., Sandor, C., Han, B., Xu, H., Stranger, B. E., Liu, X. S. *et al.* (2013). Chromatin marks identify critical cell types for fine mapping complex trait variants. *Nat Genet*, **45**, (2), 124-130.

Trynka, G., Zhernakova, A., Romanos, J., Franke, L., Hunt, K. A., Turner, G. *et al.* (2009). Coeliac disease-associated risk variants in TNFAIP3 and REL implicate altered NF-kappaB signalling. *Gut*, **58**, (8), 1078-1083.

Turner, S. and Cherry, N. (2000). Rheumatoid arthritis in workers exposed to silica in the pottery industry. *Occup Environ Med*, **57**, (7), 443-447.

Udagawa, N., Kotake, S., Kamatani, N., Takahashi, N., and Suda, T. (2002). The molecular mechanism of osteoclastogenesis in rheumatoid arthritis. *Arthritis Res*, **4**, (5), 281-289.

Untergasser, A., Cutcutache, I., Koressaar, T., Ye, J., Faircloth, B. C., Remm, M. *et al.* (2012). Primer3--new capabilities and interfaces. *Nucleic Acids Res*, **40**, (15), e115.

Valton, A. L. and Dekker, J. (2016). TAD disruption as oncogenic driver. *Curr Opin Genet Dev*, **36**, 34-40.

van Berkum, N. L., Lieberman-Aiden, E., Williams, L., Imakaev, M., Gnirke, A., Mirny, L. A. *et al.* (2010). Hi-C: a method to study the three-dimensional architecture of genomes. *J Vis Exp* (39).

van de Werken, H. J., de Vree, P. J., Splinter, E., Holwerda, S. J., Klous, P., de, W. E. *et al.* (2012a). 4C technology: protocols and data analysis. *Methods Enzymol*, **513**, 89-112.

van de Werken, H. J., Landan, G., Holwerda, S. J., Hoichman, M., Klous, P., Chachik, R. *et al.* (2012b). Robust 4C-seq data analysis to screen for regulatory DNA interactions. *Nat Methods*, **9**, (10), 969-972.

van Steensel, B. and Dekker, J. (2010). Genomics tools for unraveling chromosome architecture. *Nat Biotechnol*, **28**, (10), 1089-1095.

Vance, K. W. and Ponting, C. P. (2014). Transcriptional regulatory functions of nuclear long noncoding RNAs. *Trends Genet*, **30**, (8), 348-355.

Vaquerizas, J. M., Kummerfeld, S. K., Teichmann, S. A., and Luscombe, N. M. (2009). A census of human transcription factors: function, expression and evolution. *Nat Rev Genet*, **10**, (4), 252-263.

Venter, J. C., Adams, M. D., Myers, E. W., Li, P. W., Mural, R. J., Sutton, G. G. *et al.* (2001). The sequence of the human genome. *Science*, **291**, (5507), 1304-1351.

Verdugo, R. A., Farber, C. R., Warden, C. H., and Medrano, J. F. (2010). Serious limitations of the QTL/microarray approach for QTL gene discovery. *BMC Biol*, **8**, 96.

Vereecke, L., Beyaert, R., and van, L. G. (2011). Genetic relationships between A20/TNFAIP3, chronic inflammation and autoimmune disease. *Biochem Soc Trans*, **39**, (4), 1086-1091.

Verlaan, D. J., Ge, B., Grundberg, E., Hoberman, R., Lam, K. C., Koka, V. *et al.* (2009). Targeted screening of cis-regulatory variation in human haplotypes. *Genome Res*, **19**, (1), 118-127.

Viatte, S., Plant, D., and Raychaudhuri, S. (2013). Genetics and epigenetics of rheumatoid arthritis. *Nat Rev Rheumatol*, **9**, (3), 141-153.

Visel, A., Rubin, E. M., and Pennacchio, L. A. (2009). Genomic views of distant-acting enhancers. *Nature*, **461**, (7261), 199-205.

Wallace, C., Cutler, A. J., Pontikos, N., Pekalski, M. L., Burren, O. S., Cooper, J. D. *et al.* (2015). Dissection of a Complex Disease Susceptibility Region Using a Bayesian Stochastic Search Approach to Fine Mapping. *PLoS Genet*, **11**, (6), e1005272.

Wang, L., Song, G., Zheng, Y., Wang, D., Dong, H., Pan, J. *et al.* (2015). miR-573 is a negative regulator in the pathogenesis of rheumatoid arthritis. *Cell Mol Immunol*.

Wang, S., Wen, F., Wiley, G. B., Kinter, M. T., and Gaffney, P. M. (2013). An Enhancer Element Harboring Variants Associated with Systemic Lupus Erythematosus Engages the TNFAIP3 Promoter to Influence A20 Expression. *PLoS Genet*, **9**, (9), e1003750.

Wang, Y. Y., Wang, Q., Sun, X. H., Liu, R. Z., Shu, Y., Kanekura, T. *et al.* (2014). DNA hypermethylation of the forkhead box protein 3 (FOXP3) promoter in CD4+ T cells of patients with systemic sclerosis. *Br J Dermatol*, **171**, (1), 39-47.

Ward, L. D. and Kellis, M. (2012). HaploReg: a resource for exploring chromatin states, conservation, and regulatory motif alterations within sets of genetically linked variants. *Nucleic Acids Res*, **40**, (Database issue), D930-D934.

Ward, L. D. and Kellis, M. (2016). HaploReg v4: systematic mining of putative causal variants, cell types, regulators and target genes for human complex traits and disease. *Nucleic Acids Res*, **44**, (D1), D877-D881.

Wei, C. L., Wu, Q., Vega, V. B., Chiu, K. P., Ng, P., Zhang, T. *et al.* (2006). A global map of p53 transcription-factor binding sites in the human genome. *Cell*, **124**, (1), 207-219.

Welter, D., MacArthur, J., Morales, J., Burdett, T., Hall, P., Junkins, H. *et al.* (2014). The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. *Nucleic Acids Res*, **42**, (Database issue), D1001-D1006.

Westra, H. J. and Franke, L. (2014). From genome to function by studying eQTLs. *Biochim Biophys Acta*.

Westra, H. J., Peters, M. J., Esko, T., Yaghootkar, H., Schurmann, C., Kettunen, J. *et al.* (2013). Systematic identification of trans eQTLs as putative drivers of known disease associations. *Nat Genet*, **45**, (10), 1238-1243.

Whyte, W. A., Orlando, D. A., Hnisz, D., Abraham, B. J., Lin, C. Y., Kagey, M. H. *et al.* (2013). Master transcription factors and mediator establish super-enhancers at key cell identity genes. *Cell*, **153**, (2), 307-319.

Wingett, S., Ewels, P., Furlan-Magaril, M., Nagano, T., Schoenfelder, S., Fraser, P. *et al.* (2015). HiCUP: pipeline for mapping and processing Hi-C data. *F1000Res*, **4**, 1310.

Wright, F. A., Sullivan, P. F., Brooks, A. I., Zou, F., Sun, W., Xia, K. *et al.* (2014). Heritability and genomics of gene expression in peripheral blood. *Nat Genet*, **46**, (5), 430-437.

Wright, J. B., Brown, S. J., and Cole, M. D. (2010). Upregulation of c-MYC in cis through a large chromatin loop linked to a cancer risk-associated single-nucleotide polymorphism in colorectal cancer cells. *Mol Cell Biol*, **30**, (6), 1411-1420.

WTCCC (2007). Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature*, **447**, (7145), 661-678.

Wu, H. J., Ivanov, I. I., Darce, J., Hattori, K., Shima, T., Umesaki, Y. *et al.* (2010). Gut-residing segmented filamentous bacteria drive autoimmune arthritis via T helper 17 cells. *Immunity*, **32**, (6), 815-827.

Wu, J., Smith, L. T., Plass, C., and Huang, T. H. (2006). ChIP-chip comes of age for genome-wide functional analysis. *Cancer Res*, **66**, (14), 6899-6902.

Xia, K., Shabalin, A. A., Huang, S., Madar, V., Zhou, Y. H., Wang, W. *et al.* (2012). seeQTL: a searchable database for human eQTLs. *Bioinformatics*, **28**, (3), 451-452.

Yang, T. P., Beazley, C., Montgomery, S. B., Dimas, A. S., Gutierrez-Arcelus, M., Stranger, B. E. *et al.* (2010). Genevar: a database and Java application for the analysis and visualization of SNP-gene associations in eQTL studies. *Bioinformatics*, **26**, (19), 2474-2476.

Ye, C. J., Feng, T., Kwon, H. K., Raj, T., Wilson, M. T., Asinovski, N. *et al.* (2014). Intersection of population variation and autoimmunity genetics in human T cell activation. *Science*, **345**, (6202), 1254665.

Ye, J., Coulouris, G., Zaretskaya, I., Cutcutache, I., Rozen, S., and Madden, T. L. (2012). Primer-BLAST: a tool to design target-specific primers for polymerase chain reaction. *BMC Bioinformatics*, **13**, 134.

Yeoh, N., Burton, J. P., Suppiah, P., Reid, G., and Stebbings, S. (2013). The role of the microbiome in rheumatic diseases. *Curr Rheumatol Rep*, **15**, (3), 314.

Yochum, G. S. (2011). Multiple Wnt/ss-catenin responsive enhancers align with the MYC promoter through long-range chromatin loops. *PLoS One*, **6**, (4), e18966.

Zhang, C., Zhu, K. J., Liu, H., Quan, C., Liu, Z., Li, S. J. *et al.* (2015a). The TNFAIP3 polymorphism rs610604 both associates with the risk of psoriasis vulgaris and affects the clinical severity. *Clin Exp Dermatol*, **40**, (4), 426-430.

Zhang, J., Markus, J., Bies, J., Paul, T., and Wolff, L. (2012a). Three murine leukemia virus integration regions within 100 kilobases upstream of c-myb are proximal to the 5' regulatory region of the gene through DNA looping. *J Virol*, **86**, (19), 10524-10532.

Zhang, J., Poh, H. M., Peh, S. Q., Sia, Y. Y., Li, G., Mulawadi, F. H. *et al.* (2012b). ChIA-PET analysis of transcriptional chromatin interactions. *Methods*, **58**, (3), 289-299.

Zhang, M., Peng, L. L., Wang, Y., Wang, J. S., Liu, J., Liu, M. M. *et al.* (2016). Roles of A20 in autoimmune diseases. *Immunol Res*, **64**, (2), 337-344.

Zhang, X., Cowper-Sal, l. R., Bailey, S. D., Moore, J. H., and Lupien, M. (2012c). Integrative functional genomics identifies an enhancer looping to the SOX9 gene disrupted by the 17q24.3 prostate cancer risk locus. *Genome Res*, **22**, (8), 1437-1446.

Zhang, Y., McCord, R. P., Ho, Y. J., Lajoie, B. R., Hildebrand, D. G., Simon, A. C. *et al.* (2012d). Spatial organization of the mouse genome and its role in recurrent chromosomal translocations. *Cell*, **148**, (5), 908-921.

Zhang, Z. and Bridges, S. L., Jr. (2001). Pathogenesis of rheumatoid arthritis. Role of B lymphocytes. *Rheum Dis Clin North Am*, **27**, (2), 335-353.

Zhang, Z. and Zhang, R. (2015b). Epigenetics in autoimmune diseases: Pathogenesis and prospects for therapy. *Autoimmun Rev*, **14**, (10), 854-863.

Zhao, Z., Tavoosidana, G., Sjolinder, M., Gondor, A., Mariano, P., Wang, S. *et al.* (2006). Circular chromosome conformation capture (4C) uncovers extensive networks of epigenetically regulated intra- and interchromosomal interactions. *Nat Genet*, **38**, (11), 1341-1347.

Zheng, J., Petersen, F., and Yu, X. (2014). The role of PTPN22 in autoimmunity: learning from mice. *Autoimmun Rev*, **13**, (3), 266-271.

Zhou, Q., Long, L., Shi, G., Zhang, J., Wu, T., and Zhou, B. (2013a). Research of the methylation status of miR-124a gene promoter among rheumatoid arthritis patients. *Clin Dev Immunol*, **2013**, 524204.

Zhou, X., Lowdon, R. F., Li, D., Lawson, H. A., Madden, P. A., Costello, J. F. *et al.* (2013b). Exploring long-range genome interactions using the WashU Epigenome Browser. *Nat Methods*, **10**, (5), 375-376.

Zhou, Y., Qiu, X., Luo, Y., Yuan, J., Li, Y., Zhong, Q. *et al.* (2011). Histone modifications and methyl-CpG-binding domain protein levels at the TNFSF7 (CD70) promoter in SLE CD4+ T cells. *Lupus*, **20**, (13), 1365-1371.

8. Appendix 1

8.1. Materials

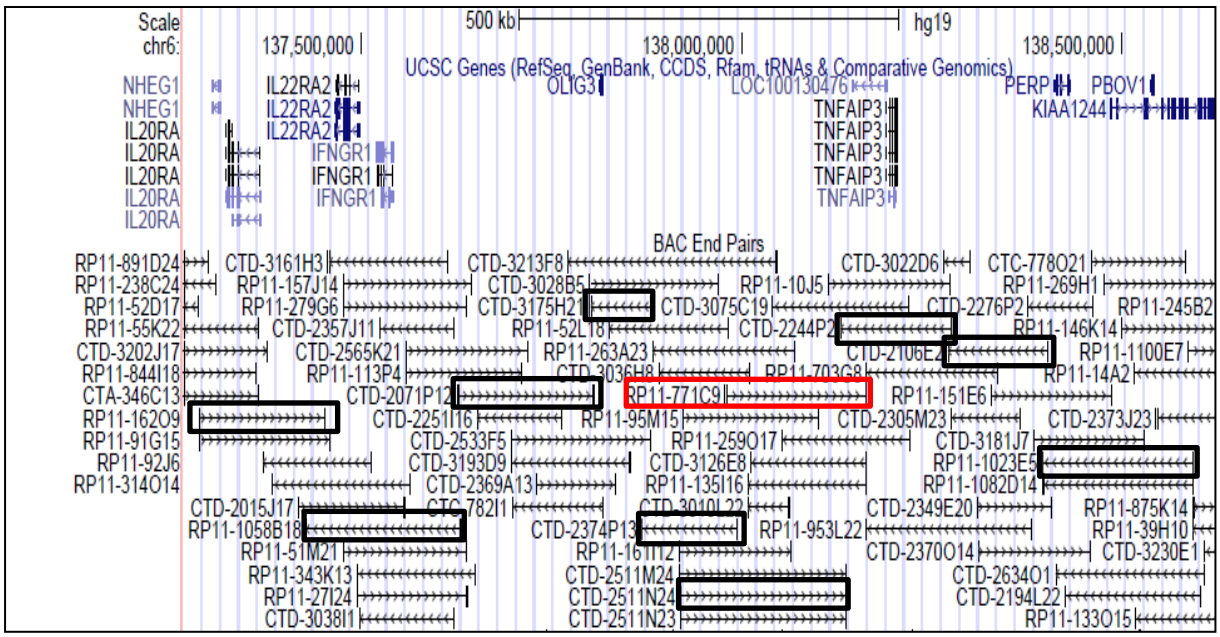
8.1.1. Cell lines

Table 24: HapMap B Lymphoblastoid cell lines (LCLs)

Cell line ID	rs6927172 Genotype	Cell line ID	rs6927172 Genotype	Cell line ID	rs6927172 Genotype
GM12878	CG	**GM10850**	GG	**GM12892**	CC
GM12865	CG	**GM10858**	GG	**GM07056**	CC
GM12875	CG	**GM12560**	GG	**GM10843**	CC
GM06993	CG			**GM10848**	CC
GM10831	CG			**GM11993**	CC
GM11994	CG			**GM06985**	CC
GM12145	CG			**GM10838**	CC
GM12812	CG			**GM10860**	CC
				GM12056	CC
				GM12707	CC

8.1.2. Bacterial artificial chromosomes (BACs)

Figure 65: UCSC track showing BAC clones spanning the 6q23 region validated using 3C-qPCR



NOTE: The RP11-771C9 BAC was incorrect when analysed by PCR and was replaced by CTD-2511N24.

Table 25: ID and co-ordinates of BAC clones used in 3C-qPCR validation of the 6q23 region

BAC clone ID	Start position	End Position	Size	Region
RP11-162O9	Chr6: 137286536	Chr6: 137450559	164024 (+ strand)	6q23
RP11-1058B18	Chr6: 137425741	Chr6: 137630056	204316 (- strand)	6q23
CTD-2071P12	Chr6: 137625360	Chr6: 137804064	178705 (+ strand)	6q23
CTD- 3175H21	Chr6: 137799448	Chr6: 137879516	80069 (- strand)	6q23
CTD- 2374P13	Chr6: 137867732	Chr6: 137992452	124721 (- strand)	6q23
CTD- 2511N24 (replacement for RP11-771C9 which was incorrect)	Chr6: 137916906	Chr6: 138137421	220516 (- strand)	6q23
CTD-2244P2	Chr6: 138128388	Chr6: 138273180	144793 (- strand)	6q23
CTD-2106E2	Chr6: 138268647	Chr6: 138400874	132228 (- strand)	6q23
RP11-1023E5	Chr6: 138393699	Chr6: 138591433	197735 (- strand)	6q23

8.1.3. Primers and adapters used in Hi-C experiments

Custom oligonucleotides for PCR and adapters for sequencing were ordered from Sigma Aldrich. All oligonucleotides were HPLC purified and supplied lyophilised. Primers and adapters were resuspended in sterile, nuclease-free water to make a 100µM stock according to the volumes specified on the supplied data sheet.

8.1.3.1. Primers used for Hi-C QC and library amplification

Table 26: Primers used in Hi-C QC PCR for short-range and long-range interactions

Primer	Sequence
HindIII_Dekker_FWD (Lieberman-Aiden *et al.* 2009)	5' – GTTCATCTTGCTGCCAGAAATGCCGAGCCTG-3'
HindIII_Dekker_REV (Lieberman-Aiden *et al.* 2009)	5' - ATCCCAGCTGTCTGTAGCTTTAGAAAGTGGG-3'
AHF64_Dekker (Belton *et al.* 2012)	5' - GCATGCATTAGCCTCTGCTGTTCTCTGAAATC-3'
AHF66_Dekker (Belton *et al.* 2012)	5' - CTGTCCAAGTACATTCCTGTTCACAAACCC-3'
6q23 region specific	
rs6927172_1_FWD	5' – TGGCCCTTAAACATAGAAAAACA-3'
rs6927172_2_REV	5' – TCCAGTTCTGGTAACCATTCTC-3'
TNFAIP3_1_FWD	5' - CTGGTCATTATGGGCTTTGG-3'
TNFAIP3_2_REV	5' - CTTCATGAATGGGGATCCAG-3'
Long range (Fraser lab):	
Human_Myc_G2	5' - GGAGAACCGGTAATGGCAAA-3'
Human_Roger_1R	5' - TGCCTGATGGATAGTGCTTTC-3'
Human_Myc_O3	5' - AAAATGCCCATTTCCTTCTCC-3'
Human_Myc_540	5' - GCATTCTGAAACCTGAATGCTC-3'

Table 27: Adapters and Primers used for Hi-C library amplification

Adapter/Primer	Sequence
TruPE_adapter_1	5'- P-GATCGGAAGAGCACACGTCTGAACTCCAGTCAC-3'
TruPE_adapter_2	5'- ACACTCTTTCCCTACACGACGCTCTTCCGATC*T-3'
Primers for pre-capture PCR	
TruPE_PCR_1.0.33	5'- ACACTCTTTCCCTACACGACGCTCTTCCGATCT-3'
TruPE_PCR_2.0.33	5'-GTGACTGGAGTTCAGACGTGTGCTCTTCCGATC-3'
Barcoded primers for final (post-capture) PCR	
TruSeq_Universal_adapter	5'- AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCT-3'
TruSeq_Index_Adapter_rc_003	5' –CAAGCAGAAGACGGCATACGAGATGCCTAAGTGACTGGAGTTCAGACGTGTGCTCTTCCGATC-3'
TruSeq_Index_Adapter_rc_006	5' –CAAGCAGAAGACGGCATACGAGATATTGGCGTGACTGGAGTTCAGACGTGTGCTCTTCCGATC-3'
TruSeq_Index_Adapter_rc_012	5' –CAAGCAGAAGACGGCATACGAGATTACAAGGTGACTGGAGTTCAGACGTGTGCTCTTCCGATC-3'
TruSeq_Index_Adapter_rc_019	5' –CAAGCAGAAGACGGCATACGAGATTTTCACGTGACTGGAGTTCAGACGTGTGCTCTTCCGAT-3'

Barcoding gives each sample a unique identity, allowing multiple pooled samples to be run on the same sequencing lane (Illumina MiSeq or HiSeq), reducing the cost of sequencing. Each sample has specific adapters ligated to the ends, one of which is Universal and contains the sequence that attaches the sample to the flow cell. The other adapter contains the barcode, which is a specific set of 6 nucleotides in the middle of the sequence used to identify the sample.

Table 28: Illumina barcoding strategy for multiplex samples

Plexity	Adapters to use
2 samples	AR005 + AR019
	AR006 + AR012
3 samples	AR002 + AR007 + AR019
	AR005 + AR006 + AR015
	2-Plex options + Any other adapter
4 samples	AR002 + AR004 + AR007 + AR016
	AR003 + AR006 + AR012 + AR019
	3-Plex options + Any other adapter

8.1.3.2. Primers used in 3C-qPCR assays

Table 29: 3C primers

Region Co-ordinates	Primer ID	Sequence	Primer co-ordinates
chr6: 137377336-137381475 (3' IL2RA)	IL20RA_1_B	TCCGAAGAGCTTTGTTTGTGG	chr6:137380925-137381497
chr6: 137403878-137407040 (3' IL2RA)	IL20RA_2_B	TGCTGCCCAGACATAGGAAA	chr6:137406409-137407071
chr6: 137421229-137423210 (3' IL2-RA)	IL20RA_3_B	AATGCGACTGTCAAGGATGC	chr6:137422283-137423263
chr6: 137570290-137583223	IFNGR1_1	CAAGGCAAGGTGGTGGTTTT	chr6:137582390-137583286
chr6: 137952897-137959707	SNPs_1_B	AGACAGACTTGAGTGCCTATTG	chr6:137958975-137959756
chr6: 137959709-137963083	SNPs_2_B	CCAGCAGGCAGAGAAAGAAT	chr6:137962245-137963208
chr6: 137983020-137989382	SNPs_3	CTGACTTTGTGATCCGCCTG	chr6:137988549-137989444
chr6: 138007203-138017056 (next fragment after P.7)	SNPs_5_B	GTCCCACCTCTGTCCAAAGA	chr6:138016473-138017091
chr6: 138025956-138036419	psPTPN11_1	ATCCCACCTGGCTGTCTATG	chr6:138035782-138036456
chr6: 138105291-138121041	Y_RNA	TCCATATCCCGTCAGCACAA	chr6:138120247-138121079
chr6: 138184709-138186854	TNFAIP3_4_B	GGCTTTGGAGTAACACAGGC	chr6:138186446-138186927
chr6: 138186856-138192635 (5' of gene)	TNFAIP3_2_B	AGCCCTCATCGACAGAAACA	chr6:138191725-138192666
chr6: 138192730-138193357 (intronic)	TNFAIP3_1_B	TCTGTGCTGTTCTGCCAATG	chr6:138192886-138193383
chr6: 138202662-138204711 (3' of gene)	TNFAIP3_3_B	AGGAAAGGGATGCTAGGACC	chr6:138204072-138204770
chr6: 138233163-138241189	lncRNA_4_B	AGTCTAGCTGGTTTGGGAGG	chr6:138240393-138241236
chr6: 138262495-138267565	lncRNA_1_B	TCAGACTGTGGAGCTTGAGG	chr6:138266740-138267611
chr6: 138267567-138268650	lncRNA_3	CTAAAGCAGACCAAGCCACC	chr6:138268060-138268686
chr6: 138320836-138334122	Down	GCCGAAATGCCTGCTATGTT	chr6:138332848-138334268
chr6:137802546-137921001	NCR _A1	ACAGGCAGTGGTATGTTGGA	chr6:137823600-137834160
chr6:137837364-137841021	NCR _B1	GGGGCTCAGTGTTCTCAGAT	chr6:137840670-137841039
chr6:137413793-137414948	P_3	ATGGTTCTGCAAGGCTGTG	chr6:137413504-137415237
chr6:137492384-137506744	P_4	GTTGTCTGCCTCTGGATCCC	chr6:137490524-137508507
chr6:137998601-138003252	P_6	CAGGTGTGAGCCATAATGCC	rs69271712 fragment chr6:137997763-138004226 (+ RA SNP rs10499194)
chr6:138003254-138007201	P_7	GAGCAGGAAATGGAGGGAGG	rs6920220 fragment chr6:138003011-138007523
chr6:138184709-138186854	P_10	GGCTTTGGAGTAACACAGGC	chr6:138184173-138187391

8.1.3.3. ChIP qPCR primers

Target Primers were designed to exclude the SNP region (X) from the design to obtain flanking primers as shown in the example below.

```
  1  AAGAAAAGCAGAAGTTTCCTGCTACTTAAGTCTAGAATGAAACAGGATTCATAATCTCTA
 61  TATTTCGGAGCTAATCAAGTGGCAATGTCAATGGGGGAAATCCAGGTCAACTTTAATTCC
           >>>>>>>>>>>>>>>>>>>>>       XXXXXXXXXX
121  CATCATGACTAAGTGGCAGAAAGACTCAGAATACAATGACCTAGAACTGACCTCAGGGAG
                                                  <<<<<<<<<<<<<<<
181  AAATCACAGTCAAATCAAGATCGCT
     <<<<<<
```

Positive control primers for the NF-κB qPCR assays were designed using an NF-κB binding site (-316/-15) in the human IkBα gene promoter. For the histone ChIP qPCR published GAPDH positive control primers were purchased from Sigma Aldrich and additional positive controls designed in regions showing high enrichment of the respective histone mark. Negative control primers were designed in gene deserts, in regions showing no evidence of transcription factor binding or histone marks.

Table 30: Primers used in ChIP-qPCR

Primer	Sequence
NF-κB	
Positive_1_F	5'- GGAATTTCCAAGCCAGTCAG – 3'
Positive_1_R	5'- CTCATCGCAGGGAGTTTCTC – 3'
Neg_1_F	5'- TGGGCAACAAGAATGAGACC – 3'
Neg_1_R	5'- TTAAAGGCAGGGGATGAGTG – 3'
Neg_2_F	5'- TGAGAATCACCAGGGCTCTT – 3'
Neg_2_R	5'- CCACTGGGGTGACTCTTAGC – 3'
H3K4me1	
POS coGAPDH_FWD (Sigma Aldrich)	5'- TAGAGGGGTGATGTGGGGAG – 3'
POS coGAPDH_REV (Sigma Aldrich)	5'- AGTGATGGCATGGACTGTGG – 3'
Neg_1_FWD	5'- CTGAAGGGATCTGGCACAGT – 3'
Neg_1_REV	5'- CCAACTGGCCAGGTAGTGAT – 3'
H3K27ac	
Neg_3_FWD	5'- ACTCATAACTCCCAGGTGCG – 3'
Neg_3_REV	5'- CGAAGGCCAAAAACTGGTCC – 3'
Pos_3_FWD	5'- CACTTGCAGAGGGACAGGAT – 3'
Pos_3_REV	5'- GAGAAACTCCCTGCGATGAG – 3'
Target	
rs6927172 Target_1_F	5'- GGAGCTAATCAAGTGGCAATG – 3'
rs6927172 Target_1_R	5'- TGATTTCTCCCTGAGGTCAGTT – 3'

8.1.4. General materials

Table 31: Equipment used in experiments

Equipment	Manufacturer
CASY automated cell counter	CASY
CO_2 cell culture incubator (Galaxy 170S)	New Brunswick Scientific
Covaris S220 instrument	Covaris
LTC2 low temperature circulator	Grant
Dell Vostro Laptop running SonoLab7	Dell
Large chilled benchtop centrifuge	Eppendorf
Microcentrifuge	Sigma
Shaking incubator	Stuart
Waterbath	Grant
Thermoblock	Eppendorf
Rotating wheel	Stuart
Magnetic rack	Life Technologies
Bioanalyzer 2100 Expert instrument	Agilent
NanoDrop 1000 instrument	Life Technologies
T100 Thermal cycler	Biorad
Veriti Thermal cycler	Applied Biosystems
QuantStudio 12k flex instrument	Life Technologies
Qubit fluorometer	Life Technologies
Vacuum concentrator	Eppendorf
MiSeq Instrument	Illumina

Table 32: General lab reagents

Material	Supplier	Cat#	Application
RPMI 1640	Sigma	R8758	Cell culture
FBS	Gibco/Life Tech	10270-106	Cell culture
Recovery cell culture freezing medium	Life technologies	12648010	Cell culture
Isopropanol	Fisher	10378923	Nucleobond preps
Triton-X 100	Sigma	T8787-50ML	Buffers
NP-40 (Igepal CA-630)	Sigma	I8890-50ML	Buffers
NEB Buffer 2	NEB	B7002S	Hi-C/3C
Formaldehyde (37% Solution)	Fisher	10532955	Hi-C/3C/ChIP
Glycine	Sigma	50046-250G	Hi-C/3C/ChIP
PBS without calcium chloride	Sigma	D5773	Hi-C/3C/ChIP
Tris-HCl pH8.0	Life Technologies	15568-025	Buffers
NaCl 5M	Promega	V4221	Buffers
Protease Inhibitor Cocktail	Roche	11873580001	Hi-C/3C/ChIP
10% SDS (Ultrapure)	Life Technologies	15553-027	Hi-C/3C/Buffers
LiCl	Sigma	L9650	ChIP buffers
Sodium Deoxycholate	Sigma	D6750	ChIP buffers
BSA	Sigma	A9418	ChIP buffers
$NaHCO_3$	Sigma	S5761	ChIP buffers
Hind III	NEB	R0104T	Hi-C/3C/BAC library
dNTPs	Life Technologies	10297-018	Hi-C/3C
Biotin-14-dATP	Life Technologies	19524-016	Hi-C
Klenow (DNA polymerase Large fragment)	NEB	M0210L	Hi-C
T4 DNA Ligase buffer	NEB	B0202S	Hi-C/3C
10mg/ml BSA	NEB	B9001S	Hi-C/3C/
T4 DNA ligase	NEB	M0202S	Hi-C/3C
T4 DNA ligase	Life Technologies	15224-025	Hi-C
Proteinase-K	Roche	03115879001	Hi-C/3C/ChIP
RNaseA	Roche	10109142001	Hi-C/3C/ChIP
Phenol	Sigma	P4557	Hi-C/3C

Table 32 continued

Material	Supplier	Cat #	Application
Phenol:Chloroform:Isoamyl Alcohol	Sigma	P3803	Hi-C/3C
Chloroform:Isoamyl Alcohol	Sigma	25666	3C control libraries
3M Sodium Acetate pH5.2	Lonza	51203	Hi-C/3C
100% Ethanol	Fisher	10437341	Hi-C/3C/ChIP
NheI HF	NEB	R3131S	Hi-C/3C
HindIII HF	NEB	R3104S	Hi-C/3C
T4 DNA Polymerase	NEB	M0203L	Hi-C/3C
0.5M EDTA	Sigma	E7889-100ML	Buffers
T4 Polynucleotide Kinase	NEB	M0201L	Hi-C
Klenow exo$^-$	NEB	M0212L	Hi-C
Agencourt Ampure XP Beads	Beckman Coulter	A63881	Hi-C
Dynabeads MyOne Streptavidin C1	Life Technologies	65001	Hi-C
Dynabeads MyOne Streptavidin T1	Life Technologies	65601	Capture Hi-C
Dynabeads Protein G	Life Technologies	10003D	ChIP
Dynabeads Protein A	Life Technologies	10001D	ChIP
Phusion Polymerase	NEB	M0530S	Hi-C/3C
Hot Star DNA Polymerase	Qiagen	203205	Hi-C/3C
MyTaq HS Polymerase	Bioline	BIO-21111	BAC identity check
SYBR green master mix	Life Technologies	4367659	ChIP/3C
MassRuler DNA ladder mix	Thermo Scientific	SM0403	Hi-C/3C/ChIP
6X MassRuler loading dye	Thermo Scientific	R0621	Hi-C/3C/ChIP
Agarose	Bioline	BIO-41025	Hi-C/3C/ChIP
Ethidium bromide	Sigma-Aldrich	E1510	Hi-C/3C/ChIP
10x TBE buffer	Thermo Scientific	10754914	Hi-C/3C/ChIP
NFkB p50	Abcam	ab7971	ChIP
NFkB p65	Abcam	ab7970	ChIP
BCL3	Santa Cruz	sc-185	ChIP
H3K4me1	Abcam	ab8895	ChIP
H3K27ac	Abcam	ab4729	ChIP

Table 33: Kits used in experiments

Kit description	Supplier	Cat#
Nucleobond BAC 100 kit	Macherey Nagel	740579
Quant-IT dsDNA broad-range kit	Life Technologies	Q33130
MinElute purification kit	Qiagen	28004
PCR purification kit	Qiagen	28104
Bioanalyzer chips – DNA HS	Agilent	5067-4626
KAPA SYBR FAST ABI Prism qPCR kit	Anachem	KK4835
SureSelectXT reagents	Agilent	G9611A
SureSelectXT Custom 0.5-2.9Mb custom DNA bait libraries	Agilent	Promoter capture design ID: 0656631 Region capture design ID: 0656641
MiSeq V3 150 cycle kit	Illumina	MS-102-3001

Table 34: General consumables

Material	Supplier	Cat#
Phase Lock Gel Tubes (Light) 50ml and 2ml	5-Prime	2ml = 2302820 50ml = 2302860
PCR strips	Agilent	410022
PCR optical strip cap	Agilent	401425
MicroTubes AFA Fiber with Snap-Cap	Covaris	Part no 520045
1ml TC12x12 AFA tubes (for chromatin shearing)	Covaris	Part no 520081
MicroAmp 384-well optical plates	Life Technologies	4309849
Microamp optical adhesive film	Life Technologies	4311971

8.2. Supplementary CHi-C data

8.2.1. SNP selection for Capture Hi-C (JIA/PsA)

Table 35: JIA and PsA SNP selection for CHi-C

Locus	Start	End	Interval size	Index SNP	No. of SNPs in LD
1p36.23	8260448	8273177	12729	rs11121129	10
1p36.11	24508746	24520350	11604	rs7552167	18
1p36.11	25289733	25305172	15439	rs77419309, **rs7523412**	54
1p31.3	67597692	67743552	145860	rs12030867, rs9988642, rs12044149	76
1p13.2	114303807	114377568	73761	rs6679677, **rs2476601**	4
1q21.3	152550017	152603842	53825	rs6693105	45
1q21.3	154291717	154428505	136788	rs4845618, **rs11265608**	61
1q24.2	167417883	167436300	18417	rs2056626	23
1q24.3	172715701	172715702	1	rs78037977	1
1q31.3	197470145	197781198	311053	rs2477077	36
2p16.1	61068821	61092678	23857	rs1306395	29
2p15	62515300	62516544	1244	rs6713082	2
2p14	65431110	65639280	208170	rs111825814	68
2q11.2	97392706	97463529	70823	rs1318597	15
2q11.2	100813330	100837567	24237	rs6740838	8
2q14.2	119566304	119577577	11273	rs11123495	8
2q24.2	162992003	163124637	132634	rs35667974	8
2q32.3	191943741	191973034	29293	rs10174238	12
2q33.2	204691537	204777818	86281	rs3087243	37
2q37.2	231175548	231187167	11619	rs10933330	4
3p21.32	44146257	44177438	31181	rs6796191	43
3p21.31	46414974	46414975	1	rs62625034	
4q27	123161618	123540758	379140	rs6849238, rs59867199	134
5p13.2	35933517	35962873	29356	rs2289878	17
5p13.1	40442868	40482599	39731	rs4957300	70
5q11.2	55436850	55442249	5399	rs71624119	5
5q15	96217691	96373750	156059	rs62376445, rs10038651	135
5q23.3	129534358	129583833	49475	rs2188958	3
5q31.1	131436216	131556174	119958	rs715285, rs7703009	32
5q31.1	131556173	131556174	1	rs17622517	1
5q31.1	131556202	131556203	1	rs7703009	1
5q31.1	131556202	131803537	247335	rs17622517, rs715285	2
5q31.1	131803536	131996669	193133	rs17622517, rs4705862, rs848	12
5q33.1	150464016	150478318	14302	rs76956521	10
5q33.3	158764176	158829527	65351	rs12188300, rs4921482	18
6p25.3	512070	513279	1209	rs7761186	4
6p22.3	20599516	20599517	1	rs112028044	1

6p22.2	25842950	26072992	230042	rs1408272	5
6q21	106560434	106560435	1	rs9320149	1
6q21	111913261	111913262	1	rs33980500	1
6q22.33	128329860	128339699	9839	rs72975941	2
6q23.3	135739354	135909796	170442	rs11154801	19
6q23.3	138185451	138234085	48634	rs610604	23
6q25.3	159323436	159324407	971	rs75402062	3
7p15.3	22765335	22810166	44831	rs1474348	41
7p15.1	28152192	28246989	94797	rs10260837	20
7p14.1	37372613	37396303	23690	rs73112675	17
7p14.1	38573225	38573234	9	rs2392581	2
8p23.1	11387860	11389224	1364	rs4841550	3
8q24.12	119886922	119932741	45819	rs2055101	8
9p21.1	32455261	32455674	413	rs1133071	2
9q31.2	109918692	109941431	22739	rs796754	2
9q32	117607259	117641001	33742	rs7048073	6
9q33.2	123640499	123721510	81011	rs7039505	84
10p15.1	6069852	6097283	27431	rs10905668, rs2025346, **rs61839660**	11
10q22.3	81067479	81067480	1	rs1972346	1
10q23.31	90759915	90782827	22912	rs1800623	18
10q26.2	129048449	129065484	17035	rs7895120	6
11p12	36336262	36437868	101606	rs12295535, rs4755450	41
11q13.1	64110421	64141771	31350	rs645078	31
11q22.3	109959637	110000795	41158	rs4561177	25
11q24.3	128502495	128504704	2209	rs4936059	8
12p13.33	260397	262646	2249	rs4980854	6
12p13.31	6506389	6519837	13448	rs7300170	3
12q13.3	56623346	56754371	131025	rs2020854	133
12q24.12	111865048	112273499	408451	rs11065991	10
12q24.13	112486817	112840766	353949	rs17630235	6
13q14.1	40299841	40368601	68760	rs7993214	27
13q14.11	42952074	43064910	112836	rs9533117	107
13q21.2	61140077	61188132	48055	rs995085	7
14q13.2	69250890	69250891	1	rs8016947	1
14q24.1	69250890	69260588	9698	rs12435329	7
16p13.13	11371758	11446647	74889	rs12928822, rs12922409	76
16p11.2	30719157	30719158	1	rs72793373	1
17q11.2	25921417	26201218	279801	rs4795067, rs8072199, rs2948521	27
17q21.2	40271756	40271757	1	rs730086	1
17q25.3	78174672	78178893	4221	rs11652075	14
18p11.21	12774325	12809340	35015	rs2847293	29
18q21.2	51777726	51843005	65279	rs602422	81
18q22.2	67524972	67528151	3179	rs34594414	2
19p13.2	10427720	10492274	64554	rs34725611	16

19p13.2	10770304	10853296	82992	rs892085	13
19p13.1	13105332	13122612	17280	rs8103241	3
19p13.11	18290414	18340910	50496	rs62120394	22
19q13.42	55766805	55791091	24286	rs12983085	6
20q13.13	48514173	48590791	76618	rs6063454	47
20q13.33	62336257	62373707	37450	rs4809330	27
21q22.12	36695908	36745167	49259	rs9979383	6
21q22.3	45635061	45652514	17453	rs2298565	17
22q11.21	21911219	21983260	72041	rs2298428, rs5749502	95
22q12.1	37531116	37636351	105235	rs2284033	44

8.2.2. Combined list of loci for Capture Hi-C

Table 36: Combined list of RA, PsA and JIA loci included in the CHi-C design

Chromosome	Start	End	Size bp	Disease associated SNP
chr1	2,483,960	2,751,364	267,404	rs2843401|rs187786174
chr1	7,956,773	7,969,309	12,536	rs227163
chr1	8,260,448	8,273,177	12,729	rs11121129
chr1	17,644,462	17,676,172	31,710	rs2301888|rs2240336
chr1	24,508,746	24,520,347	11,601	rs7552167
chr1	25,289,733	25,305,172	15,439	rs77419309|rs7523412
chr1	38,260,502	38,362,803	102,301	rs28411352
chr1	38,614,866	38,644,861	29,995	rs12140275|rs883220
chr1	67,597,692	67,658,954	61,262	rs12044149|rs12030867
chr1	67,684,933	67,743,552	58,619	rs9988642
chr1	114,303,807	114,377,568	73,761	rs2476601|chr1:114303808-114377568|rs2476601|rs2476601|rs6679677
chr1	117,259,268	117,280,696	21,428	rs624988|rs798000
chr1	152,550,017	152,603,842	53,825	rs6693105
chr1	154,395,124	154,428,505	33,381	rs4845618|rs2228145
chr1	154,291,717	154,379,369	87,652	rs11265608
chr1	157,668,992	157,705,725	36,733	rs2317230
chr1	160,831,047	160,831,048	1	rs4656942
chr1	161,399,919	161,450,597	50,678	rs72717009
chr1	161,463,875	161,483,977	20,102	rs10494360
chr1	161,644,257	161,847,068	202,811	rs75409195
chr1	167,417,883	167,434,277	16,394	rs2056626
chr1	172,668,339	172,793,418	125,079	chr1:172668340-172793418|rs78037977
chr1	173,306,645	173,353,881	47,236	rs2105325
chr1	197,470,145	197,781,198	311,053	rs2477077
chr1	198,598,662	198,670,555	71,893	rs17668708
chr1	200,780,152	200,832,857	52,705	chr1:200780153-200832857

chr1	206,939,903	206,957,449	17,546	chr1:206939904-206957449
chr2	30,443,459	30,449,594	6,135	rs10175798
chr2	61,068,821	61,167,216	98,395	rs1306395|rs34695944|rs34695944
chr2	62,437,099	62,464,950	27,851	rs13385025
chr2	62,515,300	62,516,544	1,244	rs6713082
chr2	65,431,110	65,518,430	87,320	rs111825814
chr2	65,545,067	65,635,872	90,805	rs6546146|rs1858037|rs11689314
chr2	97,392,706	97,463,529	70,823	rs1318597
chr2	100,636,756	100,870,862	234,106	rs10209110|chr2:100658077-100870862|rs9653442|rs6740838
chr2	111,601,477	111,616,141	14,664	rs6732565
chr2	119,566,304	119,577,577	11,273	rs11123495
chr2	162,992,003	163,237,390	245,387	rs35667974|rs2111485|chr2:163110536-163237390
chr2	191,900,448	191,973,034	72,586	rs13426947|rs11889341|rs10174238
chr2	202,151,491	202,193,463	41,972	rs6715284
chr2	204,586,514	204,649,276	62,762	rs1980422|rs1980422
chr2	204,691,537	204,781,918	90,381	chr2:204691538-204745003|rs3087243|rs3087243|rs11571302
chr2	231,175,548	231,187,167	11,619	rs10933330
chr3	16,985,271	17,076,560	91,289	rs4452313
chr3	27,758,273	27,793,632	35,359	rs3806624
chr3	44,146,963	44,177,438	30,475	rs6796191
chr3	46,150,936	46,541,541	390,605	chr3:46150937-46541541|rs62625034
chr3	58,183,635	58,318,477	134,842	rs35677470|rs73081554
chr3	136,159,127	136,632,122	472,995	rs9826828
chr4	10,727,356	10,727,357	1	rs13142500
chr4	26,031,094	26,134,258	103,164	chr4:26031095-26134258|rs11933540|rs932036
chr4	48,220,838	48,220,839	1	rs2664035
chr4	79,493,842	79,513,215	19,373	rs10028001
chr4	123,073,008	123,540,758	467,750	rs6849238
chr4	166,573,266	166,575,267	2,001	chr4:166573267-166575267
chr5	35,800,546	35,927,309	126,763	chr5:35800547-35927309
chr5	35,933,517	35,962,873	29,356	rs2289878
chr5	40,442,868	40,482,599	39,731	rs4957300
chr5	55,444,682	55,444,683	1	rs7731626
chr5	55,436,850	55,444,640	7,790	rs71624119|rs71624119|chr5:55438580-55444640
chr5	96,117,862	96,203,033	85,171	rs62376445
chr5	96,220,086	96,373,750	153,664	rs10038651
chr5	102,595,777	102,686,157	90,380	rs39984|rs2561477
chr5	129,534,358	129,583,833	49,475	rs2188958
chr5	131,357,410	131,556,203	198,793	rs657075|rs7703009|rs715285

chr5	131,803,536	131,803,537	1	rs17622517
chr5	131,813,218	131,832,514	19,296	rs4705862
chr5	131,995,842	131,996,669	827	rs848
chr5	150,464,578	150,478,318	13,740	rs76956521
chr5	158,764,176	158,804,928	40,752	rs4921482
chr5	158,829,526	158,829,527	1	rs12188300
chr6	426,154	426,268	114	rs9378815
chr6	512,070	513,175	1,105	rs7761186
chr6	14,087,483	14,129,630	42,147	rs74984480
chr6	20,597,454	20,711,693	114,239	rs112028044
chr6	25,715,656	26,093,141	377,485	rs1408272
chr6	32,428,771	32,428,772	1	rs9268839
chr6	36,345,839	36,358,289	12,450	rs2234067
chr6	44,228,814	44,284,899	56,085	rs2233424
chr6	90,850,163	91,012,867	162,704	chr6:90850164-91012867
chr6	106,430,084	106,508,640	78,556	rs6911690
chr6	106,560,434	106,560,435	1	rs9320149
chr6	106,629,689	106,787,169	157,480	rs9372120
chr6	111,913,261	111,913,262	1	rs33980500
chr6	126,659,042	126,903,011	243,969	chr6:126659043-126903011
chr6	128,329,860	128,339,699	9,839	rs72975941
chr6	135,739,354	135,909,796	170,442	rs11154801
chr6	137,959,234	138,006,504	47,270	rs17264332|rs6920220
chr6	138,125,880	138,243,739	117,859	rs7752903|rs610604
chr6	149,810,194	149,886,122	75,928	rs9373594
chr6	159,442,800	159,539,485	96,685	rs629326|rs2451258
chr6	159,323,436	159,324,407	971	rs75402062
chr6	167,523,394	167,546,504	23,110	rs1571878|rs59466457
chr7	22,766,220	22,810,166	43,946	rs1474348
chr7	28,152,192	28,246,989	94,797	rs10260837|rs67250450
chr7	37,372,613	37,396,303	23,690	rs73112675
chr7	38,573,225	38,573,234	9	rs2392581
chr7	92,236,163	92,237,533	1,370	rs4272
chr7	128,573,966	128,581,835	7,869	rs3778753|rs3807306
chr8	11,333,352	11,353,110	19,758	rs4840565|rs2736337
chr8	11,387,860	11,389,224	1,364	rs4841550
chr8	81,095,394	81,134,484	39,090	rs998731
chr8	102,451,262	102,469,182	17,920	rs678347
chr8	119,886,922	119,932,741	45,819	rs2055101
chr8	129,540,463	129,571,140	30,677	rs1516971
chr9	4,282,535	4,296,430	13,895	chr9:4282536-4296430
chr9	32,455,261	32,455,674	413	rs1133071
chr9	34,707,372	34,755,359	47,987	rs11574914|rs2812378
chr9	109,918,692	109,918,693	1	rs796754

chr9	117,607,259	117,693,173	85,914	rs7048073
chr9	123,636,120	123,723,351	87,231	rs10985070|rs10739580|rs7039505
chr10	6,060,432	6,063,319	2,887	chr10:6060433-6063319
chr10	6,065,941	6,067,941	2,000	chr10:6065942-6067941
chr10	6,069,852	6,070,675	823	rs2025346
chr10	6,078,552	6,176,166	97,614	rs61839660|chr10:6088743-6176166|rs10905668|rs706778|rs10795791
chr10	6,390,191	6,404,700	14,509	rs947474|rs947474
chr10	8,077,698	8,108,592	30,894	rs2275806|chr10:8095340-8108592|rs3824660
chr10	9,043,456	9,049,253	5,797	rs12413578
chr10	31,411,112	31,422,671	11,559	rs793108
chr10	50,097,818	50,097,819	1	rs2671692
chr10	63,779,870	63,813,790	33,920	rs71508903|rs12764378
chr10	64,036,880	64,044,448	7,568	rs6479800
chr10	81,065,200	81,067,480	2,280	rs1972346
chr10	81,703,372	81,758,075	54,703	rs726288
chr10	90,023,032	90,051,035	28,003	chr10:90023033-90051035
chr10	90,759,915	90,782,827	22,912	rs1800623
chr10	129,047,726	129,065,484	17,758	rs7895120
chr11	2,181,223	2,183,224	2,001	chr11:2181224-2183224
chr11	2,193,596	2,198,665	5,069	chr11:2193597-2198665
chr11	36,336,262	36,376,021	39,759	rs4755450
chr11	36,418,688	36,437,868	19,180	rs12295535
chr11	36,451,313	36,530,644	79,331	rs570676|rs331463
chr11	60,888,000	60,925,215	37,215	rs508970|rs595158
chr11	61,547,067	61,618,169	71,102	rs968567
chr11	64,097,232	64,141,771	44,539	rs645078|rs645078
chr11	72,411,663	72,416,325	4,662	rs11605042
chr11	95,311,259	95,320,808	9,549	rs4409785
chr11	107,877,138	107,970,987	93,849	rs138193887
chr11	109,959,637	110,000,795	41,158	rs4561177
chr11	118,610,548	118,745,884	135,336	rs4938573|rs10790268
chr11	128,496,951	128,496,952	1	rs73013527
chr11	128,502,495	128,504,704	2,209	rs4936059
chr12	260,869	262,646	1,777	rs4980854
chr12	6,495,274	6,519,837	24,563	rs7300170
chr12	9,824,137	9,929,679	105,542	chr12:9824138-9929679
chr12	56,379,059	56,394,954	15,895	rs773125
chr12	56,403,576	56,482,180	78,604	chr12:56403577-56482180
chr12	56,627,299	56,754,371	127,072	rs2020854
chr12	58,017,700	58,142,854	125,154	rs10683701|rs1633360
chr12	111,833,787	112,007,756	173,969	rs10774624|chr12:111884608-112007756

chr12	112,072,423	112,610,714	538,291	rs11065991|rs17630235
chr13	40,262,759	40,368,601	105,842	rs7993214|rs9603616
chr13	42,952,074	43,064,910	112,836	rs9533117
chr13	61,140,077	61,188,132	48,055	rs995085
chr13	100,078,914	100,081,766	2,852	chr13:100078915-100081766
chr14	35,832,665	35,832,666	1	rs8016947
chr14	61,908,918	62,002,703	93,785	rs3783782
chr14	68,728,424	68,760,141	31,717	rs1950897|chr14:68752593-68754593
chr14	69,250,890	69,260,588	9,698	rs12435329
chr14	98,485,110	98,498,951	13,841	chr14:98485111-98498951
chr14	101,300,566	101,307,703	7,137	chr14:101300567-101307703
chr14	105,392,836	105,392,837	1	rs2582532
chr15	38,820,646	38,920,825	100,179	rs8043085|rs8032939|chr15:38836777-38920825
chr15	69,984,461	70,010,647	26,186	rs8026898|rs8026898
chr15	79,233,713	79,235,713	2,000	chr15:79233714-79235713
chr16	11,164,566	11,238,991	74,425	chr16:11164567-11238991
chr16	11,371,758	11,446,480	74,722	rs12928822|rs12922409|chr16:11371759-11446480
chr16	11,793,394	11,841,539	48,145	rs4780401
chr16	28,490,516	28,601,186	110,670	chr16:28490517-28601186
chr16	30,719,157	30,719,158	1	rs72793373
chr16	75,236,762	75,252,327	15,565	chr16:75236763-75252327
chr16	86,005,837	86,021,627	15,790	rs13330176|rs13330176
chr17	5,136,760	5,272,580	135,820	rs72634030
chr17	25,887,570	25,921,418	33,848	rs2948521
chr17	26,106,674	26,117,407	10,733	rs4795067|rs8072199
chr17	26,182,887	26,201,218	18,331	rs9907633
chr17	37,493,597	37,740,789	247,192	rs1877030
chr17	37,908,866	38,111,419	202,553	rs12936409|chr17:37908867-38111419|rs59716545
chr17	38,753,549	38,861,757	108,208	chr17:38753550-38861757
chr17	40,271,756	40,271,757	1	rs730086
chr17	78,174,672	78,178,893	4,221	rs11652075
chr18	12,774,325	12,809,340	35,015	rs2847293|chr18:12774894-12809340
chr18	12,821,902	12,881,361	59,459	rs8083786
chr18	51,777,726	51,843,005	65,279	rs602422
chr18	67,511,644	67,546,842	35,198	chr18:67511645-67543688|rs2469434|rs34594414
chr19	10,416,443	10,590,508	174,065	chr19:10416444-10590508|rs34536443|rs34536443|rs34536443|rs34725611
chr19	10,771,940	10,771,941	1	rs147622113
chr19	10,811,666	10,853,296	41,630	rs892085

chr19	13,111,373	13,122,612	11,239	rs8103241
chr19	18,294,922	18,338,709	43,787	rs62120394
chr19	49,206,107	49,218,060	11,953	chr19:49206108-49218060
chr19	55,766,805	55,791,091	24,286	rs12983085
chr20	1,609,951	1,674,340	64,389	chr20:1609952-1674340
chr20	44,730,244	44,749,251	19,007	rs4239702|rs6032662
chr20	48,514,173	48,590,791	76,618	rs6063454
chr20	62,336,257	62,372,706	36,449	rs4809330
chr21	34,748,356	34,764,288	15,932	rs73194058
chr21	35,909,624	35,938,968	29,344	rs147868091|rs2834512
chr21	36,712,587	36,738,242	25,655	rs8133843|rs9979383|rs9979383
chr21	43,825,356	43,836,186	10,830	chr21:43825357-43836186
chr21	43,855,066	43,855,067	1	rs1893592
chr21	45,635,061	45,652,756	17,695	rs2298565|rs2236668
chr22	21,912,215	21,984,379	72,164	rs5749502|rs2298428|rs11089637
chr22	30,203,598	30,592,487	388,889	chr22:30203599-30592487
chr22	37,531,116	37,537,514	6,398	rs2284033
chr22	37,544,244	37,552,894	8,650	rs3218251|rs3218251
chr22	37,581,484	37,609,342	27,858	chr22:37581485-37609342
chr22	37,624,998	37,636,351	11,353	rs8135343
chr22	39,739,186	39,747,780	8,594	rs909685
chrX	78,464,615	78,464,616	1	rs201408742
chrX	153,195,392	153,378,375	182,983	rs5987194

chr7	27090828	27303174	HOXA - Control region
chr16	104265	263351	HBA - Control region

8.2.3. Library quality control

Table 37: Quantification of Hi-C and 3C libraries

Sample	Quantity (ng/µl)	Amount needed for 5µg biotin removal (Hi-C only)	Number of aliquots used for biotin removal
GM12878_3C_1	645	x	x
GM12878_HiC_1	825	6µl	8 (40µg)
Jurkat_3C_1	625	x	X
Jurkat_HiC_2	725	7µl	8 (40µg)
GM12878_3C_2	710	x	x
GM12878_HiC_2	875	5µl	8 (40µg)
Jurkat_3C_2	800	x	x
Jurkat_HiC_2	850	7µl	8 (40µg)

Table 38: Post size-selection quantification and number of aliquots used for streptavidin-biotin pulldown

Sample	Quantity (ng/µl)	Quantity in 190µl (µg)	Number of aliquots for pulldown
1) GM12878_HiC_Capture	43.4	8.5	4 x 2-2.5µg
2) Jurkat_HiC_Capture	53.8	10.5	5 x 2-2.5µg
5) GM12878_HiC_BRCAP	36.5	7	3 x 2-2.5µg
6) Jurkat_HiC_BRCAP	17.4	2.5	2 x 2-2.5µg

Table 39: Final amplification PCRs and sample identification

Sample	No of final amplification PCRs	Volume recovered	Final samples
1) GM12878_HiC_Capture	60 (Split 30 + 30)	A 800µl B 750µl	1) GM_ProCap 2) GM_RegCap
2) Jurkat_HiC_Capture	76 (Split 38 + 38)	A 960µl B 960µl	1) JK_ProCap 2) JK_RegCap
5) GM12878_HiC_BR	50 (Split 25 + 25)	A 595µl B 570µl	1) GM_ProCap_BR 2) GM_RegCap_BR
6) Jurkat_HiC_BR	30 (Split 15 + 15)	A 350µl B 350µl	1) JK_ProCap_BR 2) JK_RegCap_BR

8.2.4. Sample preparation for Illumina sequencing

Table 40: Preparation of diluted samples for MiSeq analysis

MiSeq Run 1					
Component	**V(f) µl**	**C(i) nM**	**C(f) nM**	**# Indexes**	**Vol (µl)**
GM_ProCap	20	17.94	4	4	1.114827
JK_ProCap	20	12.21	4	4	1.638002
GM_RegCap	20	15.71	4	4	1.273074
JK_RegCap	20	11.62	4	4	1.72117
Low TE buffer					14.25293
Final volume					20
MiSeq Run 2					
Component	**V(f) ul**	**C(i) nM**	**C(f) nM**	**# Indexes**	**Vol (ul)**
GM_ProCap_BR	20	14.16	4	4	1.412429
JK_ProCap_BR	20	11.07	4	4	1.806685
GM_RegCap_BR	20	12.94	4	4	1.545595
JK_RegCap_BR	20	35.76	4	4	0.559284
Low TE buffer					14.67601
Final volume					20

Table 41: Preparation of diluted samples for HiSeq 2500 sequencing

Pooling for HiSeq (Run 1)					
Component	**V(f) µl**	**C(i) nM**	**C(f) nM**	**# Indexes**	**Vol (µl)**
GM_ProCap	20	17.94	10	1	11.14827
Low TE buffer					8.851728
Final volume					20
Component	**V(f) µl**	**C(i) nM**	**C(f) nM**	**# Indexes**	**Vol (µl)**
JK_ProCap	20	12.21	10	1	16.38002
Low TE buffer					3.619984
Final volume					20
Component	**V(f) µl**	**C(i) nM**	**C(f) nM**	**# Indexes**	**Vol (µl)**
GM_RegCap	20	15.71	10	2	6.365372
JK_RegCap	20	11.62	10	2	8.605852
Low TE buffer					5.028776
Final volume					20

Table 42: Preparation of diluted samples for HiSeq 2500 sequencing

Pooling for HiSeq (Run 2)					
Component	**V(f) µl**	**C(i) nM**	**C(f) nM**	**# Indexes**	**Vol (µl)**
GM_ProCap_BR	20	14.16	10	1	14.12429
Low TE buffer					5.875706
Final volume					20
Component	**V(f) µl**	**C(i) nM**	**C(f) nM**	**# Indexes**	**Vol (µl)**
JK_ProCap_BR	20	11.07	10	1	18.06685
Low TE buffer					1.933153
Final volume					20
Component	**V(f) µl**	**C(i) nM**	**C(f) nM**	**# Indexes**	**Vol (µl)**
GM_RegCap_BR	20	12.94	10	2	7.727975
JK_RegCap_BR	20	35.76	10	2	2.796421
Low TE buffer					9.475604
Final volume					10

8.2.5. Sequencing QC

Table 43: Sequencing QC statistics from Illumina MiSeq QC

Cell line	Reads processed	Aligned pairs	Valid di-tags	Invalid di-tags	Unique di-tags	Cis	Trans
GM_ProCap_MiSeq_1	5,749,805	4,847,890	3,581,640	598,492	3,556,772	2,969,195	587,577
GM_RegCap_MiSeq_1	5,282,643	4,526,916	3,388,935	556,919	3,359,022	2,843,703	515,319
JK_ProCap_MiSeq_1	6,066,117	5,133,397	3,836,795	594,974	3,807,752	3,159,186	648,566
JK_RegCap_MiSeq_1	5,499,581	4,778,672	3,643,202	550,614	3,599,402	3,020,441	578,961
GM_ProCap_MiSeq_2	6,941,069	5,989,351	4,501,770	700,793	4,424,754	3,542,151	882,603
GM_RegCap_MiSeq_2	7,578,056	6,834,142	5,333,477	756,887	5,067,329	4,147,842	919,487
JK_ProCap_MiSeq_2	5,777,237	4,866,516	3,777,024	389,403	3,730,624	3,020,203	709,721
JK_RegCap_MiSeq_2	9,291,991	7,967,785	6,270,543	637,588	6,039,066	4,933,458	1,105,608
Average	**6,523,312**	**5,618,084**	4,291,673	598,209	**4,198,090**	3,454,522	743,480

Table 44: Percentages generated from MiSeq QC stats used to generate excel charts

Cell line	% aligned	% Valid di-tags	% Invalid di-tags	% Unique di-tags	% Cis	% Trans
GM_ProCap_MiSeq_1	84.3	73.9	12.3	99.3	83.5	16.5
GM_RegCap_MiSeq_1	85.7	74.9	12.3	99.1	84.7	15.3
JK_ProCap_MiSeq_1	84.6	74.7	11.6	99.2	83.0	17.0
JK_RegCap_MiSeq_1	86.9	76.2	11.5	98.8	83.9	16.1
GM_ProCap_MiSeq_2	86.3	75.2	11.7	98.3	80.1	19.9
GM_RegCap_MiSeq_2	90.2	78.0	11.1	95.0	81.9	18.1
JK_ProCap_MiSeq_2	84.2	77.6	8.0	98.8	81.0	19.0
JK_RegCap_MiSeq_2	85.7	78.7	8.0	96.3	81.7	18.3
Average	**86.0**	**76.2**	**10.8**	**98.1**	**82.4**	**17.6**

Table 45: Types of invalid di-tag (MiSeq)

Cell line	Valid_pairs	Circularised	Same_fragment_Internal	Same_fragment_Dangling	Re-ligation	Contiguous_sequence	Wrong_size	Invalid pairs
GM_ProCap_MiSeq_1	3,581,640	50,686	102,670	45,065	221,029	16,861	162,181	598,492
GM_RegCap_MiSeq_1	3,388,935	44,897	112,129	41,442	184,595	14,102	159,754	556,919
JK_ProCap_MiSeq_1	3,836,795	53,835	94,602	51,864	209,587	15,487	169,599	594,974
JK_RegCap_MiSeq_1	3,643,202	48,873	99,029	48,302	175,778	13,437	165,195	550,614
GM_ProCap_MiSeq_2	4,501,770	61,367	61,736	27,625	283,485	20,961	245,619	700,793
GM_RegCap_MiSeq_2	5,333,477	65,613	69,540	29,703	289,439	21,099	281,493	756,887
JK_ProCap_MiSeq_2	3,777,024	40,543	32,383	23,748	124,554	10,210	157,965	389,403
JK_RegCap_MiSeq_2	6,270,543	65,815	58,680	37,831	188,399	15,222	271,641	637,588
Average	**4,291,673**	**53,954**	**78,846**	**38,198**	**209,608**	**15,922**	**201,681**	**598,209**

Table 46: Percentages generated from MiSeq QC stats used to generate excel charts

Cell line	Valid_pairs	Circularised	Same_fragment_Internal	Same_fragment_Dangling	Re-ligation	Contiguous_sequence	Wrong_size	Invalid pairs
GM_ProCap_MiSeq_1	87.6	1.4	2.9	1.3	6.2	0.5	4.5	16.7
GM_RegCap_MiSeq_1	87.4	1.3	3.3	1.2	5.4	0.4	4.7	16.4
JK_ProCap_MiSeq_1	88.3	1.4	2.5	1.4	5.5	0.4	4.4	15.5
JK_RegCap_MiSeq_1	87.7	1.3	2.7	1.3	4.8	0.4	4.5	15.1
GM_ProCap_MiSeq_2	87.1	1.4	1.4	0.6	6.3	0.5	5.5	15.6
GM_RegCap_MiSeq_2	86.5	1.2	1.3	0.6	5.4	0.4	5.3	14.2
JK_ProCap_MiSeq_2	92.1	1.1	0.9	0.6	3.3	0.3	4.2	10.3
JK_RegCap_MiSeq_2	91.8	1.0	0.9	0.6	3.0	0.2	4.3	10.2
Average	**88.6**	**1.3**	**1.8**	**0.9**	**4.9**	**0.4**	**4.7**	**13.9**

Figure 66: MiSeq quality summary charts

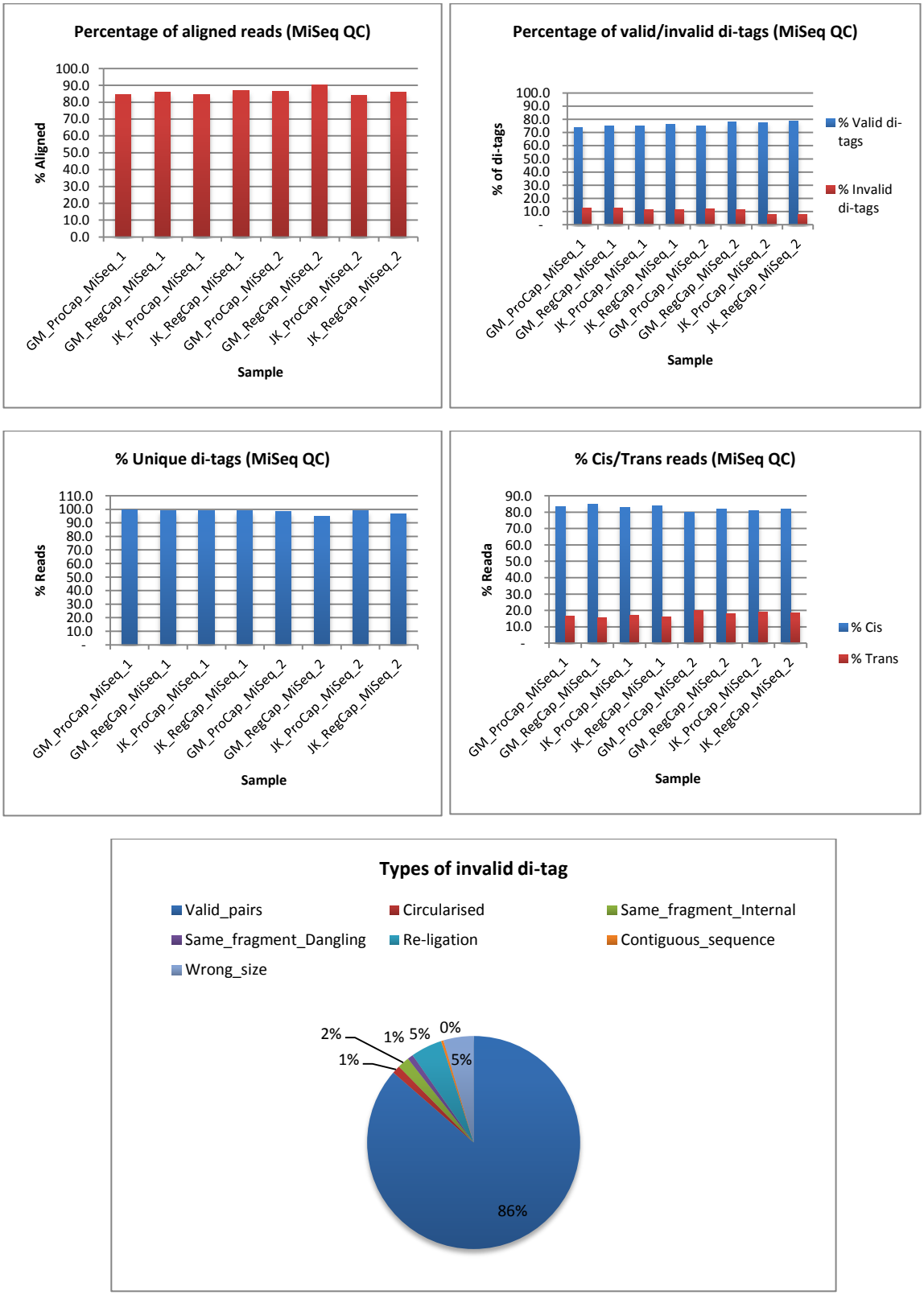


Table 47: Sequencing QC statistics from Illumina HiSeq

Cell line and experiment	Reads processed	Aligned pairs	Valid di-tags	Invalid di-tags	Unique di-tags	Cis	Trans	On-target	Bait-to-bait
GM12878_RegCap_BR1	78,467,719	59,929,732	51,465,358	8,464,374	45,892,434	38,853,120	7,039,314	20,113,771	3,359,680
GM12878_RegCap_BR2	108,735,810	88,841,743	77,796,593	11,045,150	46,255,464	37,792,853	8,462,611	40,793,883	6,199,018
Jurkat_RegCap_BR_1	76,963,038	60,126,486	52,225,839	7,900,647	44,810,487	37,562,253	7,248,234	28,073,973	5,877,743
Jurkat_RegCap_BR_2	101,765,348	77,022,176	69,953,227	7,068,949	50,891,861	41,540,063	9,351,798	26,829,286	3,173,044
GM12878_ProCap_BR1	148,618,189	110,497,620	94,604,772	15,892,848	81,500,754	68,047,789	13,452,965	46,738,753	4,917,795
GM12878_ProCap_BR2	181,179,872	137,965,848	119,415,709	18,550,139	85,847,090	68,580,262	17,266,828	74,373,047	7,411,636
Jurkat_ProCap_BR_1	155,743,979	116,474,859	100,827,324	14,539,636	85,111,648	70,572,012	14,539,636	59,494,411	4,173,962
Jurkat_ProCap_BR_2	172,658,827	126,626,485	114,899,145	11,727,340	86,965,933	70,383,387	16,582,546	55,491,937	5,073,930

Table 48: Percentages generated from HiSeq QC stats used to generate excel charts

Cell line	% aligned	% Valid di-tags	% Invalid di-tags	% Unique di-tags	% Cis	% Trans	% On-target	% Bait-to-bait
GM12878_RegCap_BR1	76.4	85.9	14.1	89.2	84.7	**15.3**	**43.8**	16.7
GM12878_RegCap_BR2	81.7	87.6	12.4	59.5	81.7	**18.3**	**88.2**	15.2
Jurkat_RegCap_BR_1	78.1	86.9	13.1	85.8	83.8	**16.2**	**62.7**	20.9
Jurkat_RegCap_BR_2	75.7	90.8	9.2	72.8	81.6	**18.4**	**52.7**	11.8
GM12878_ProCap_BR1	74.3	85.6	14.4	86.1	83.5	**16.5**	**57.3**	10.5
GM12878_ProCap_BR2	76.1	86.6	13.4	71.9	79.9	**20.1**	**86.6**	10.0
Jurkat_ProCap_BR_1	74.8	86.6	12.5	84.4	82.9	**17.1**	**69.9**	7.0
Jurkat_ProCap_BR_2	73.3	90.7	9.3	75.7	80.9	**19.1**	**63.8**	9.1
Average	**76.3**	**87.6**	**12.3**	**78.2**	**82.4**	**17.6**	**65.6**	**12.7**

Table 49: Types of invalid di-tag

Cell line	Valid_pairs	Circularised	Same_fragment_ Internal	Same_fragment_ Dangling	Re-ligation	Contiguous_ sequence	Wrong_size	Invalid pairs
GM12878_RegCap_BR1	51,465,358	675,077	625,966	1,708,127	2,785,791	216,755	2,452,658	8,464,374
GM12878_RegCap_BR2	77,796,593	959,881	1,007,642	423,950	4,189,913	311,195	4,152,569	11,045,150
Jurkat_RegCap_BR_1	52,225,839	699,769	1,429,367	691,836	2,504,038	192,503	2,383,134	7,900,647
Jurkat_RegCap_BR_2	69,953,227	736,785	646,118	417,080	2,081,981	172,094	3,014,891	7,068,949
GM12878_ProCap_BR1	94,604,772	1,334,956	2,730,530	1,187,289	5,833,143	439,143	4,367,787	15,892,848
GM12878_ProCap_BR2	119,415,709	1,622,548	1,621,602	724,813	7,484,474	550,995	6,545,707	18,550,139
Jurkat_ProCap_BR_1	100,827,324	1,408,529	2,496,040	1,370,337	5,482,984	403,333	4,486,312	15,647,535
Jurkat_ProCap_BR_2	114,899,145	1,224,092	980,741	707,243	3,753,038	307,303	4,754,923	11,727,340
Averages GM_RegCap	**64,630,976**	**817,479**	**816,804**	**1,066,039**	**3,487,852**	**263,975**	**3,302,614**	**8,619,780**
Averages_JK_RegCap	**61,089,533**	**718,277**	**1,037,743**	**554,458**	**2,293,010**	**182,299**	**2,699,013**	**7,484,798**
Averages GM_ProCap	**107,010,241**	**1,478,752**	**2,176,066**	**956,051**	**6,658,809**	**495,069**	**5,456,747**	**15,454,466**
Averages JK_ProCap	**107,863,235**	**1,316,311**	**1,738,391**	**1,038,790**	**4,618,011**	**355,318**	**4,620,618**	**13,687,438**

Table 50: Percentages generated from HiSeq QC stats used to generate excel charts

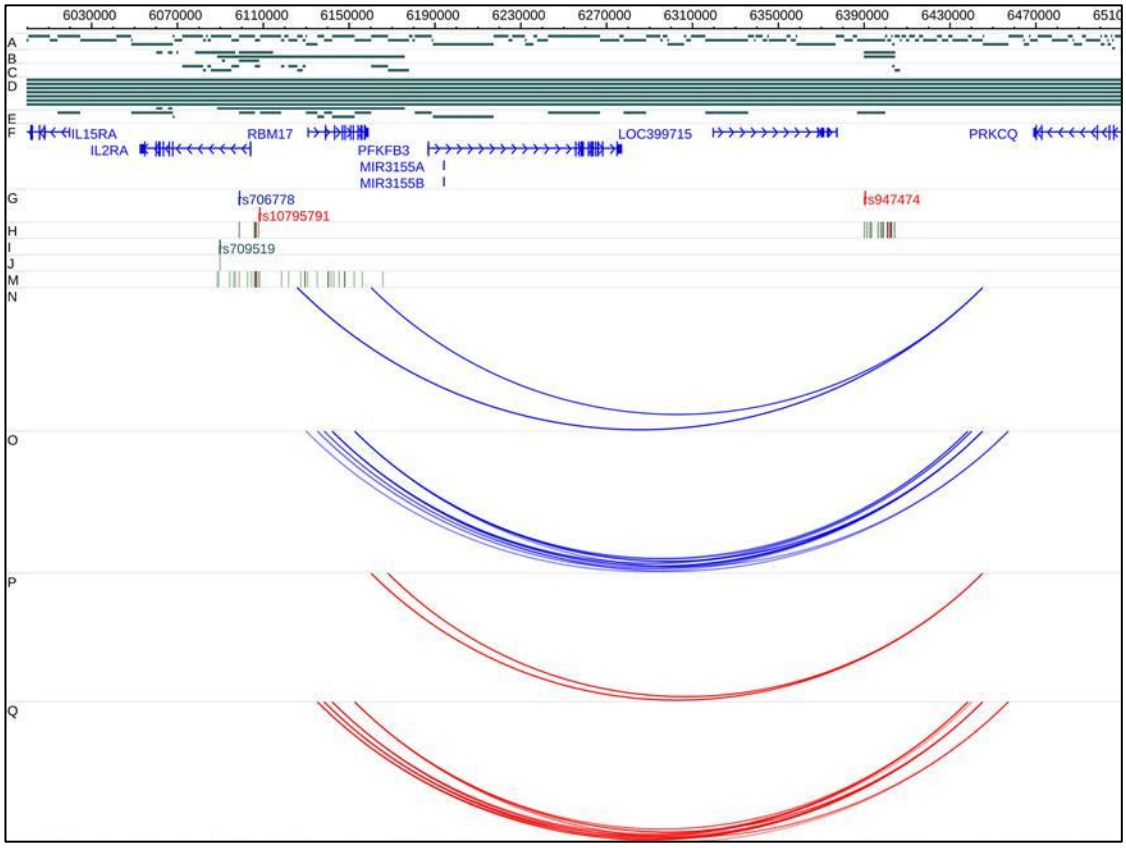
Cell line	% Valid_pairs	% Circularised	% Same_fragment_Internal	% Same_fragment_ Dangling	% Re-ligation	% Contiguous_ sequence	% Wrong_ size	% Invalid pairs
GM12878_RegCap_BR1	85.9	1.3	1.2	3.3	5.4	0.4	4.8	14.1
GM12878_RegCap_BR2	87.6	1.2	1.3	0.5	5.4	0.4	5.3	12.4
Jurkat_RegCap_BR_1	86.9	1.3	2.7	1.3	4.8	0.4	4.6	13.1
Jurkat_RegCap_BR_2	90.8	1.1	0.9	0.6	3.0	0.2	4.3	9.2
GM12878_ProCap_BR1	85.6	1.4	2.9	1.3	6.2	0.5	4.6	14.4
GM12878_ProCap_BR2	86.6	1.4	1.4	0.6	6.3	0.5	5.5	13.4
Jurkat_ProCap_BR_1	86.6	1.4	2.5	1.4	5.4	0.4	4.4	13.4
Jurkat_ProCap_BR_2	90.7	1.1	0.9	0.6	3.3	0.3	4.1	9.3
Average	87.6	1.3	1.7	1.2	5.0	0.4	4.7	12.4

8.3. Supplementary CHi-C data

Table 51: Analysis of CHi-C interactions in chromosomes 3-6

Dataset	Region	Index SNP	Potential Interactions for follow-up
Plenge_102_Promoter **Plenge_102_Region**	3p24.1	rs3806624	EOMES – AZI2
iChip_RA_ Promoter **iChip_RA_ Region** **Plenge_102_Promoter** **Plenge_102_Region** **JIA/PSA_ Region**	4p15.2	rs932036 rs11933540 rs932036	SNP region – RBPJ (GM) RBPJ - STIM2 (GM and JK)
iChip_RA_ Promoter **iChip_RA_ Region** **Plenge_102_Promoter** **Plenge_102_Region** **JIA/PSA_ Region**	5q11.2	rs71624116	ANKRD55 – IL6ST (GM, JK) ANKRD55 – DDX4 (GM, JK)
iChip_RA_ Promoter **Plenge_102_Promoter** **Plenge_102_Region** **JIA/PSA_ Promoter**	6q23	rs6920220 rs7752903, rs17264332 rs610604	TNFAIP3 – RP11-240M16.1 SNP - TNFAIP3 (GM, JK) SNP - IFNGR SNP - IL20RA SNP - RP11-240M16.1 IL22RA – IFNGR TNFAIP3 –lncRNAs, IL22RA lncRNA – lncRNA
iChip_RA_ Region **Plenge_102_Promoter** **Plenge_102_Region**	6q27	rs59466457 rs1571878	SNP – FGFR1OP FGFR1OP – CCR6
Plenge_102_Region **JIA/PSA_ Promoter**	6p25.3	rs9378815 rs7761186	IRF4 – EXOC2, FOXF2
JIA/PSA_ Region	6q25.3	rs75402062	TAGAP – SYTL3

Figure 67: Long-range interactions between *PRKCQ* and *IL2RA*



Genomic co-ordinates are shown along the top. (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; (G) Index SNPs identified for RA and (I) JIA; Density plots showing 1000 Genomes SNPs in LD (r^2>0.8) with the index SNPs for RA (H) and JIA (J); (N–Q) Significant Interactions identified in the region and promoter capture experiments in GM12878 (N,O) and Jurkat (P,Q) cells.

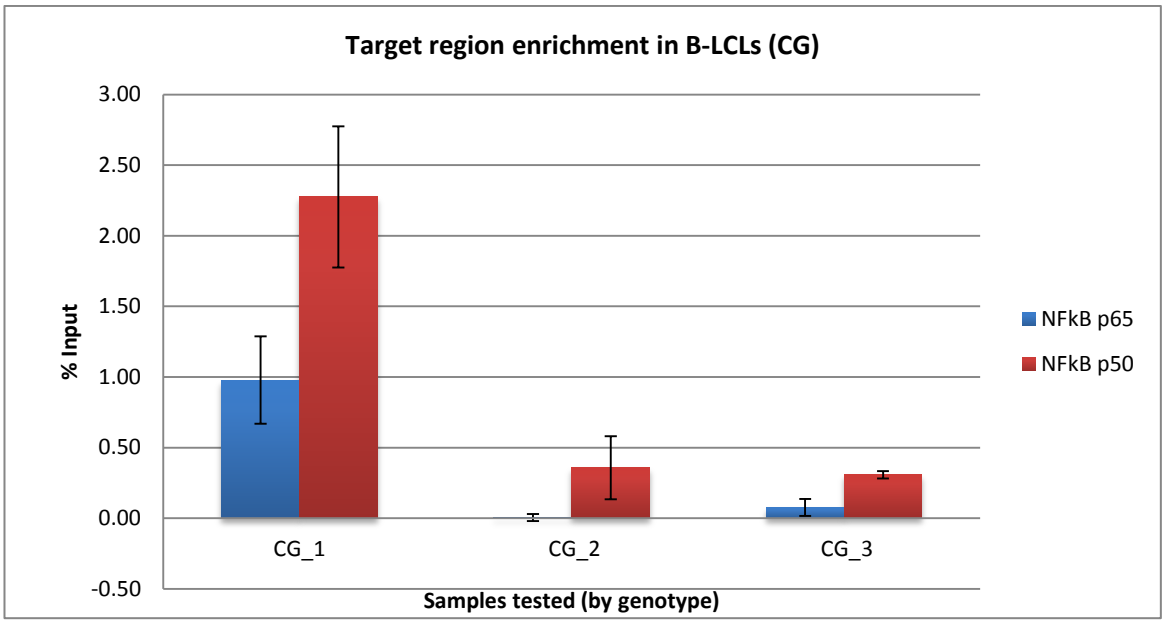
Table 52: Co-ordinates of significant interactions identified in the 6q23 locus

Left Co-ordinates	ID	Right Co-ordinates	ID
chr6: 137403878-137407040	IL20RA_2	chr6: 137952897-137959707	SNPs_1
chr6: 137570290-137583223	IFNGR1_1	chr6: 137952897-137959707	SNPs_1
chr6: 137570290-137583223	IFNGR1_1	chr6: 137959709-137963083	SNPs_2
chr6: 137983020-137989382	SNPs_3	chr6: 138262495-138267565	lncRNA_1
chr6: 138007203-138017056	SNPs_5	chr6: 138262495-138267565	lncRNA_1
chr6: 137983020-137989382	SNPs_3	chr6: 138267567-138268650	lncRNA_3
chr6: 137983020-137989382	SNPs_3	chr6: 138267567-138268650	DOWN
chr6: 138025956-138036419	psPTPN11_1	chr6: 138192730-138193357	TNFAIP3_1
chr6: 138192730-138193357	TNFAIP3_1	chr6: 138262495-138267565	lncRNA_1
chr6: 138202662-138204711	TNFAIP3_3	chr6: 138262495-138267565	lncRNA_1
chr6: 138202662-138204711	TNFAIP3_3	chr6: 138267567-138268650	lncRNA_3
chr6: 137421229-137423210	IL20RA_3	chr6: 138233163-138241189	lncRNA_4
chr6: 137421229-137423210	IL20RA_3	chr6: 138262495-138267565	lncRNA_1
chr6: 137421229-137423210	IL20RA_3	chr6: 138186856-138192635	TNFAIP3_2
chr6: 138184709-138186854	TNFAIP3_4	chr6: 138105291-138121041	Y_RNA

8.4. Supplementary ChIP data

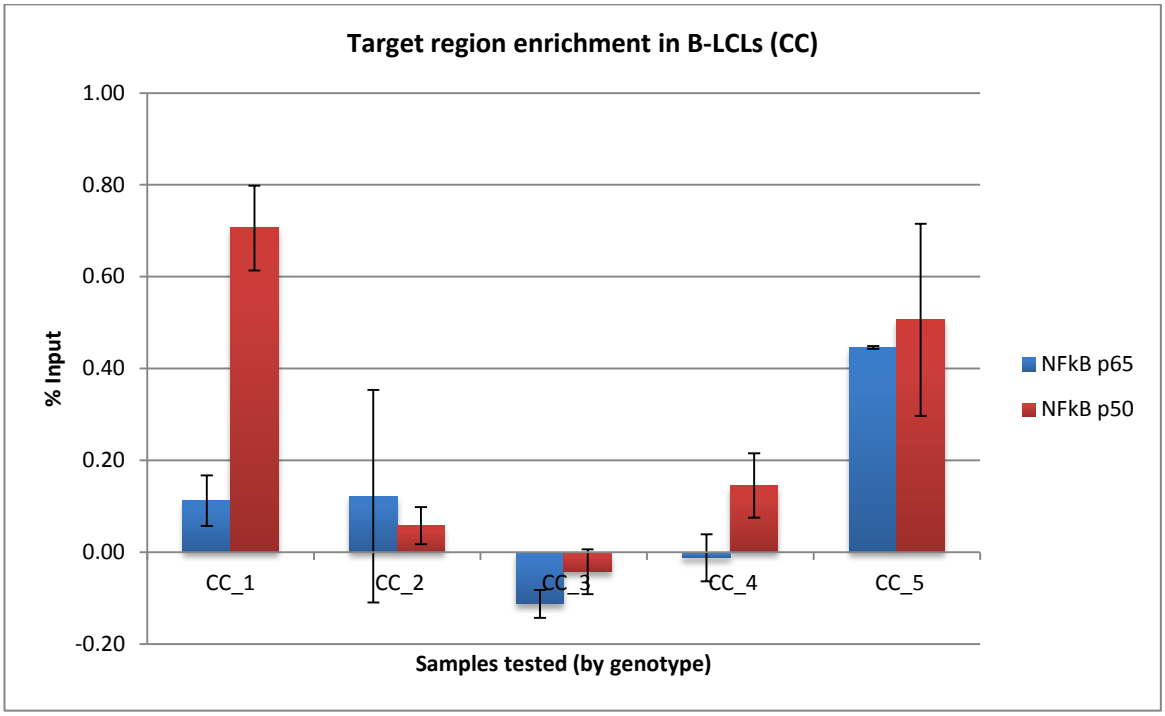
8.4.1. Figures showing variation in genotypes in ChIP assays

Figure 68: **NF-kB binding in B-cells containing rs6927172 CG genotype**



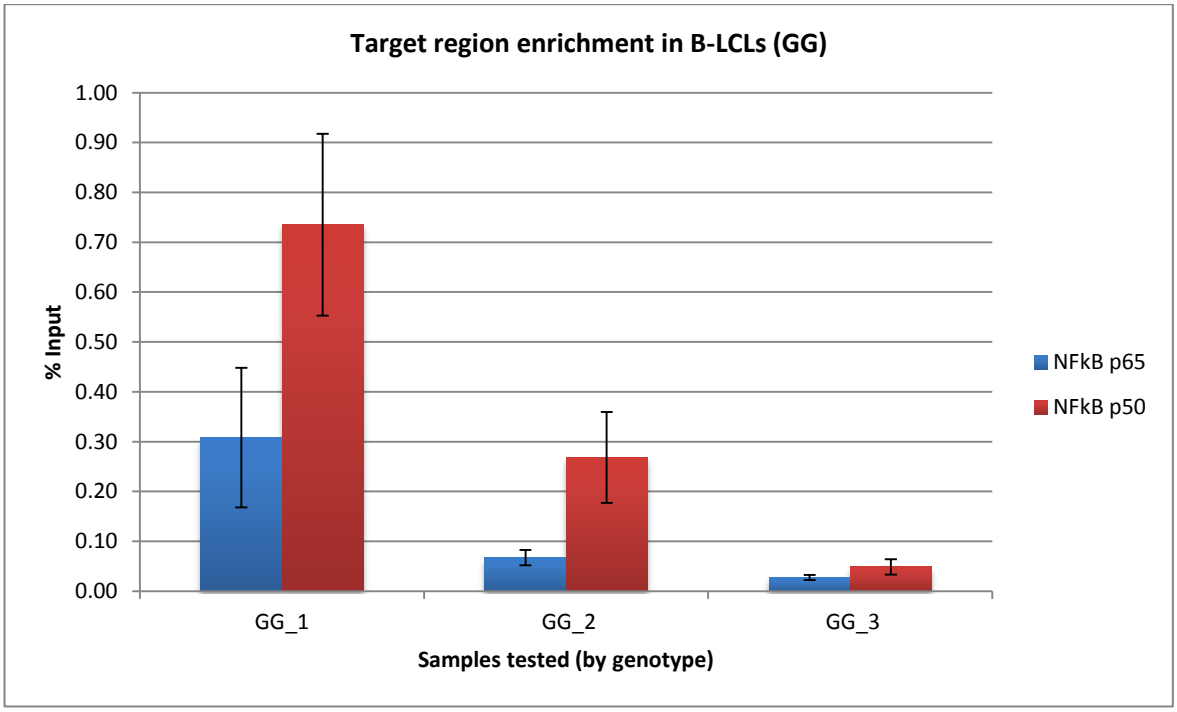
B-lymphoblastoid cell lines containing the CG rs6927172 genotype were tested for NF-kB p50 and p65 transcription factor binding at the rs6927172 target region. Each ChIP was carried out in triplicate along with a no antibody control. SYBR green qPCR was carried out in triplicate on a QuantStudio 12K Flex instrument using primers specific for the target region, a positive control region and negative control region (data not shown). ChIP samples were normalised to the non-IP'd Input sample and the no-antibody control was used to determine the level of non-specific, background binding which was subtracted off the sample values. CG_1 = GM12878, CG_2 = GM12875, CG_3 = GM12865. The data represents the average % Input results from the three samples; error bars are +/- SEM.

Figure 69: NF-kB binding in B-cells containing rs6927172 CC genotype



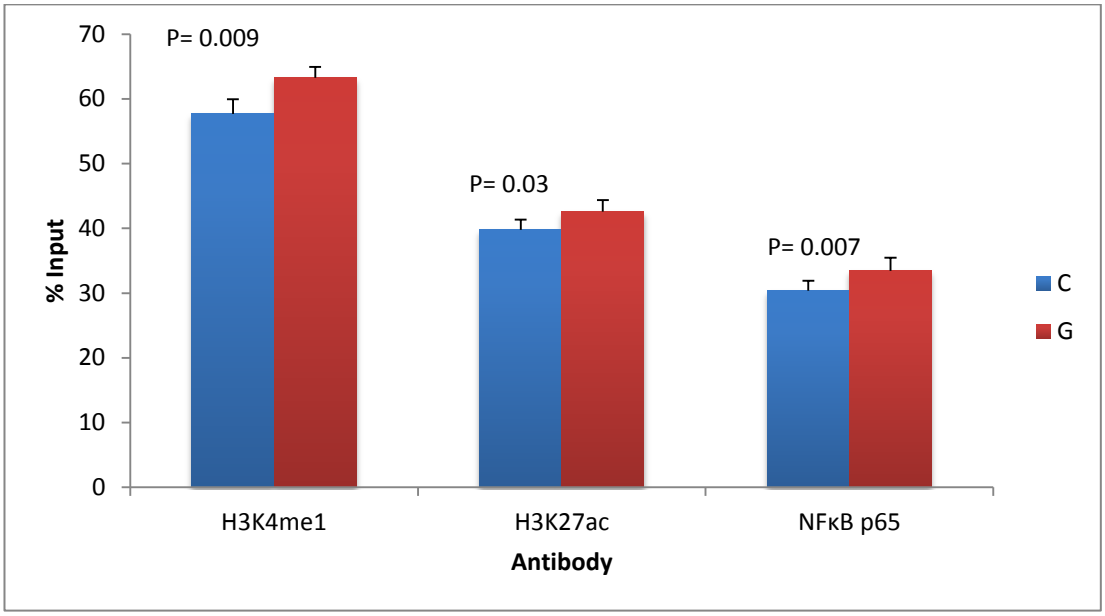
B-lymphoblastoid cell lines containing the CC rs6927172 genotype were tested for NF-kB p50 and p65 transcription factor binding at the rs6927172 target region. Each ChIP was carried out in triplicate along with a no antibody control. SYBR green qPCR was carried out in triplicate on a QuantStudio 12K Flex instrument using primers specific for the target region, a positive control region and negative control region (data not shown). ChIP samples were normalised to the non-IP'd Input sample and the no-antibody control was used to determine the level of non-specific, background binding which was subtracted off the sample values. CC_1 = GM12892, CC_2 = GM07056, CC_3 = GM10843, CC_4 = GM10848, CC_5 = GM11993. The data represents the average % Input results from the five samples; error bars are +/- SEM.

Figure 70: NF-kB binding in B-cells containing rs6927172 GG genotype



B-lymphoblastoid cell lines containing the GG (risk) rs6927172 genotype were tested for NF-kB p50 and p65 transcription factor binding at the rs6927172 target region. Each ChIP was carried out in triplicate along with a no antibody control. SYBR green qPCR was carried out in triplicate on a QuantStudio 12K Flex instrument using primers specific for the target region, a positive control region and negative control region (data not shown). ChIP samples were normalised to the non-IP'd Input sample and the no-antibody control was used to determine the level of non-specific, background binding which was subtracted off the sample values. GG_1 = GM10850, GG_2 = GM10858, GG_3 = GM12560. The data represents the average % Input results from the three samples; error bars are +/- SEM.

Figure 71: Allele specific ChIP in Jurkat cells (TaqMan)

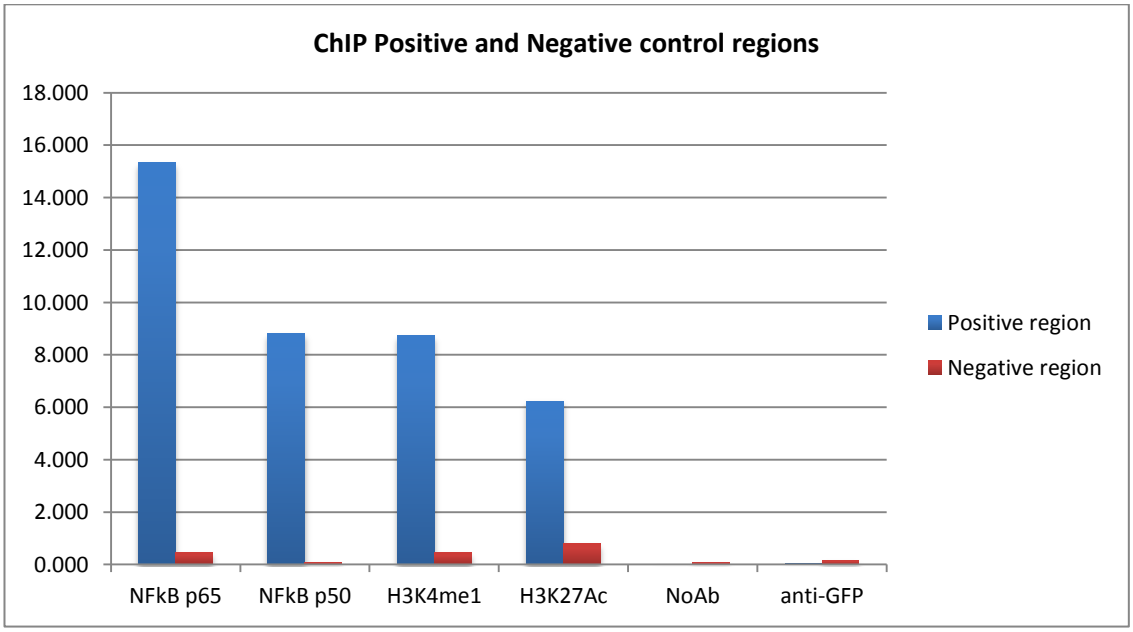


(Figure courtesy of Dr Gisela Orozco who performed the assay and data analysis for this experiment)

Enrichment of the histone marks H3K4me1 and H3K27ac, and the transcription factor NF-κB p65 at the rs6927172 target region was assessed in Jurkats using a TaqMan assay to allow for detection of allele-specific differences. Each ChIP was carried out in triplicate along with a no antibody control. TaqMan qPCR was carried out in triplicate on a QuantStudio 12K Flex instrument using probes specific for the target region. ChIP samples were normalised to the non-IP'd Input sample. The data shown represents the average % Input results from the samples tested; error bars are +/- SEM.

8.4.2. Positive and Negative Control regions

Figure 72: ChIP positive and negative control regions



8.4.3. ChIP summary data

Table 53: Normalised average target region enrichment for NF-κB antibodies (B-cells)

Normalised average data - p65	% Target region enrichment	Average	St Dev	Normalised average data - p50	% Target region enrichment	Average	St Dev
CG Genotype							
GM12878	0.98	0.3531	0.54	GM12878	2.27	0.9798	1.12
GM12875	0.00			GM12875	0.36		
GM12865	0.08			GM12865	0.31		
CC Genotype							
GM12892	0.11	0.1451	0.24	GM12892	0.71	0.2381	0.31
GM07056	0.46			GM07056	0.06		
GM10843	-0.13			GM10843	-0.08		
GM10848	-0.01			GM10848	0.15		
GM11993	0.30			GM11993	0.36		
GG Genotype							
GM10850	0.31	0.1342	0.15	GM10850	0.74	0.3509	0.35
GM10858	0.07			GM10858	0.27		
GM12560	0.03			GM12560	0.05		

Table 54: Normalised average target region enrichment for NF-κB and histone mark antibodies (T-cells)

% Input	NF-kB p65	NF-kB p50	BCL3	H3K4me1	H3K27ac	NoAb	anti-GFP
Jurkat_1	0.164	0.276	0.895	2.892	0.137	0.010	0.001
Jurkat_2	0.160	0.219	0.902	2.957	0.212	0.003	0.020
Jurkat_3	0.114	0.219	1.021	2.882	0.201	0.009	0.021
Average	**0.146**	**0.238**	**0.939**	**2.910**	**0.183**	**0.007**	**0.014**
St.Dev	**0.028**	**0.033**	**0.071**	**0.041**	**0.041**	**0.004**	**0.011**

Table 55: Normalised average target region enrichment for histone mark antibodies (B-cells)

H3K4me1	rs6927172_CC	rs6927172_CG	rs6927172_GG
	14.442	6.954	5.628
	17.016	8.160	4.015
	15.943	6.187	7.322
Average	15.800	7.100	5.655
St Dev	1.293	0.995	1.654
No_Ab	0.125	1.395	0.179

H3K27ac	rs6927172_CC	rs6927172_CG	rs6927172_GG
	2.001	2.188	1.856
	1.783	1.649	2.056
	2.616	1.828	1.574
Average	2.133	1.889	1.829
St Dev	0.432	0.066	0.340
No_Ab	0.125	0.275	0.179

Table 56: Normalised average target region enrichment for NF-κB and BCL3 antibodies (B-cells)

rs6927172	**NFkB p65**	**NFkB p50**	**BCL3**	**No_Ab**	**anti-GFP**
GM12878_1	1.187	2.698	2.793		
GM12878_2	1.374	2.793	3.574		
GM12878_3	1.265	2.751	3.156		
Average	1.275	2.747	3.174	0.060	0.046268981
No_Ab	0.060	0.060	0.060		
St Dev	0.132	0.067	0.552		

9. Appendix 2

Manuscript

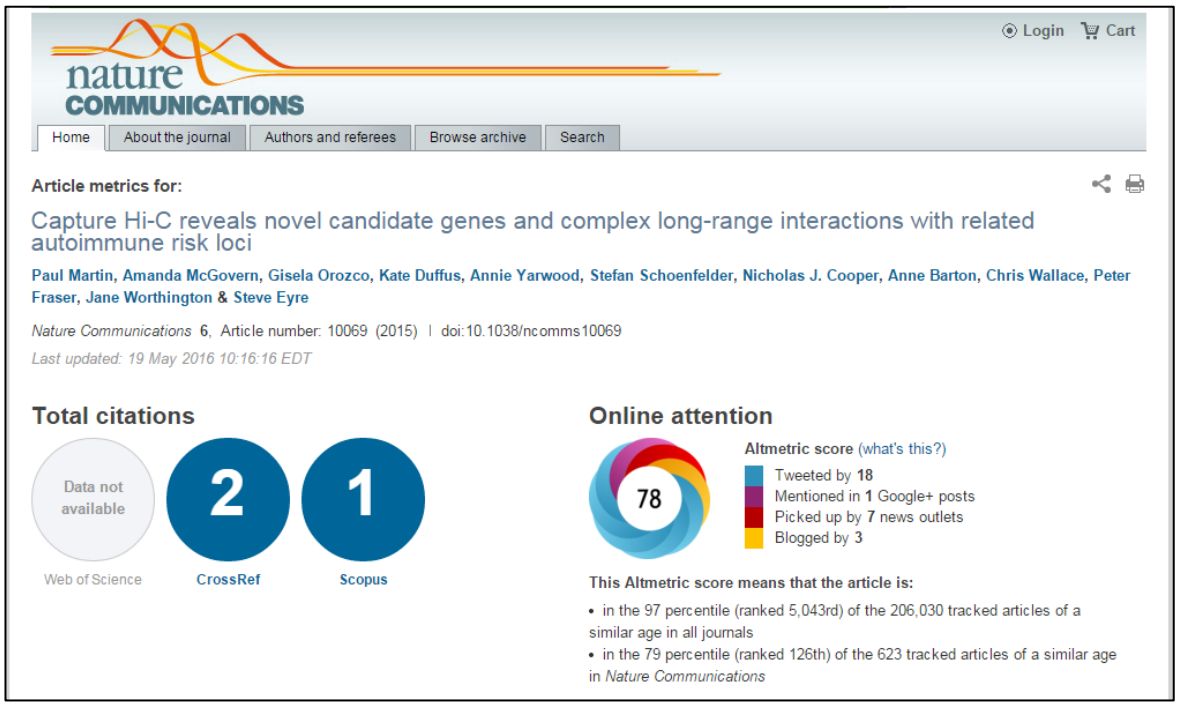
Martin, P., McGovern, A., Orozco, G., Duffus, K., Yarwood, A., Schoenfelder, S. *et al.* (2015). Capture Hi-C reveals novel candidate genes and complex long-range interactions with related autoimmune risk loci. *Nat Commun*, **6**, 10069.

http://www.nature.com/ncomms/2015/151130/ncomms10069/full/ncomms10069.html

Abstract

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).

http://www.nature.com/ncomms/2015/151130/ncomms10069/metrics



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

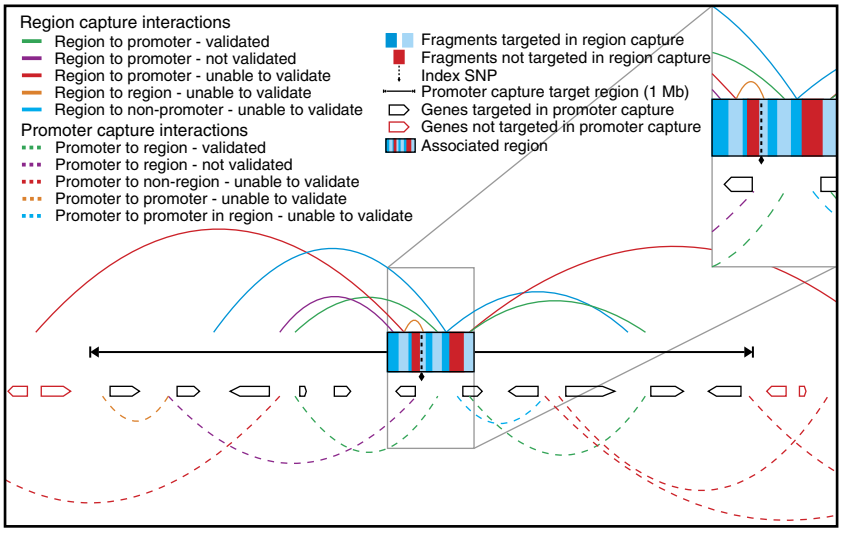


Figure 1 | A schematic of a hypothetical associated region including possible chromatin interactions. Chromatin interactions are shown by arcs, those above the promoter capture target region are observed in the 'Region Capture' experiment; those below are observed in the 'Promoter Capture' experiment. 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'.

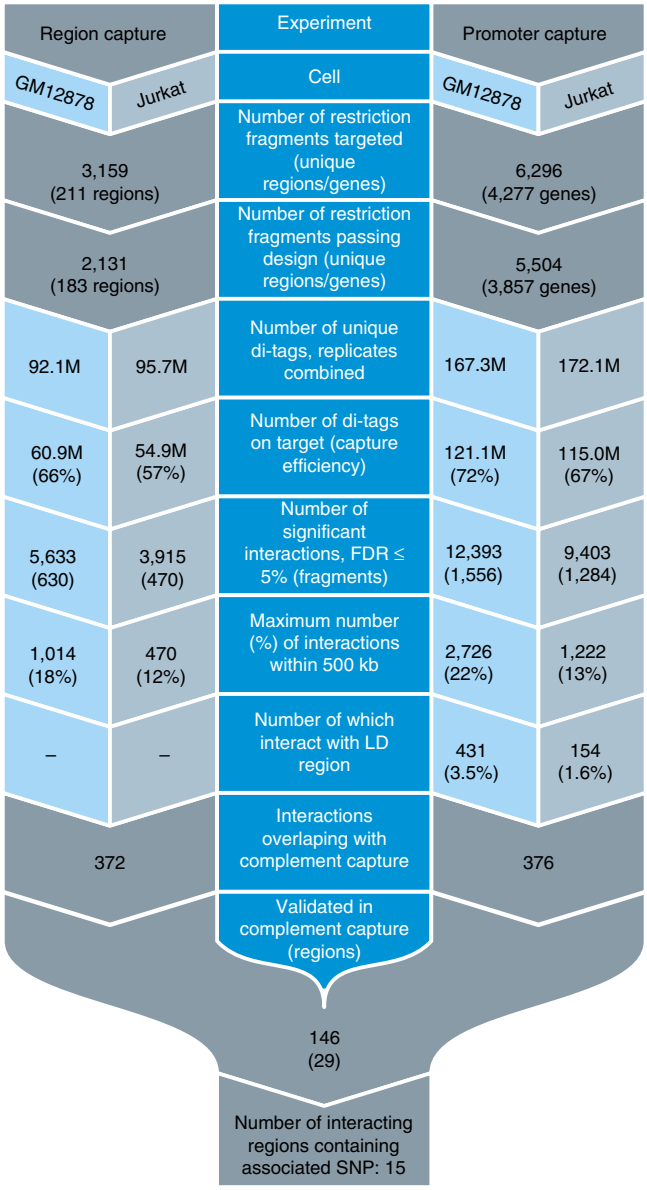


Figure 2 | Flowchart summarizing capture Hi-C experiments by cell line. The 'Region Capture' experiment is shown on the left and the 'Promoter Capture' experiment on the right. Flowchart sections are coloured by cell type: light blue—GM12878 cells; light grey—Jurkat cells and grey—both cell types. Each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

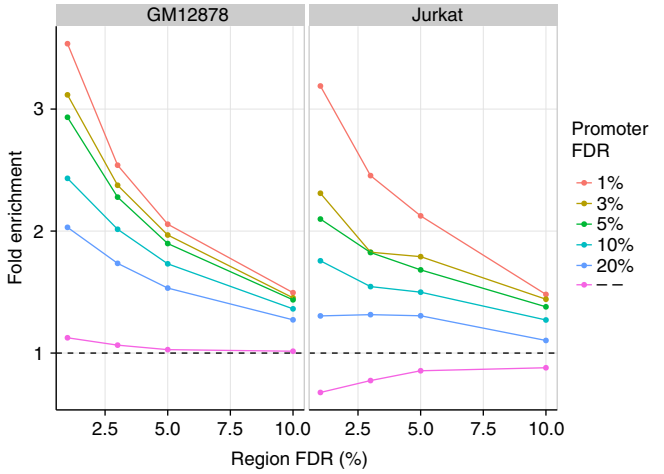


Figure 3 | Fold enrichment. Fold enrichment of retained interactions called in the promoter capture experiments with decreasing FDR thresholds, given they had been called in the region capture experiments at the FDR threshold shown. '—' shows the enrichment found by focusing only on interactions called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

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).

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*

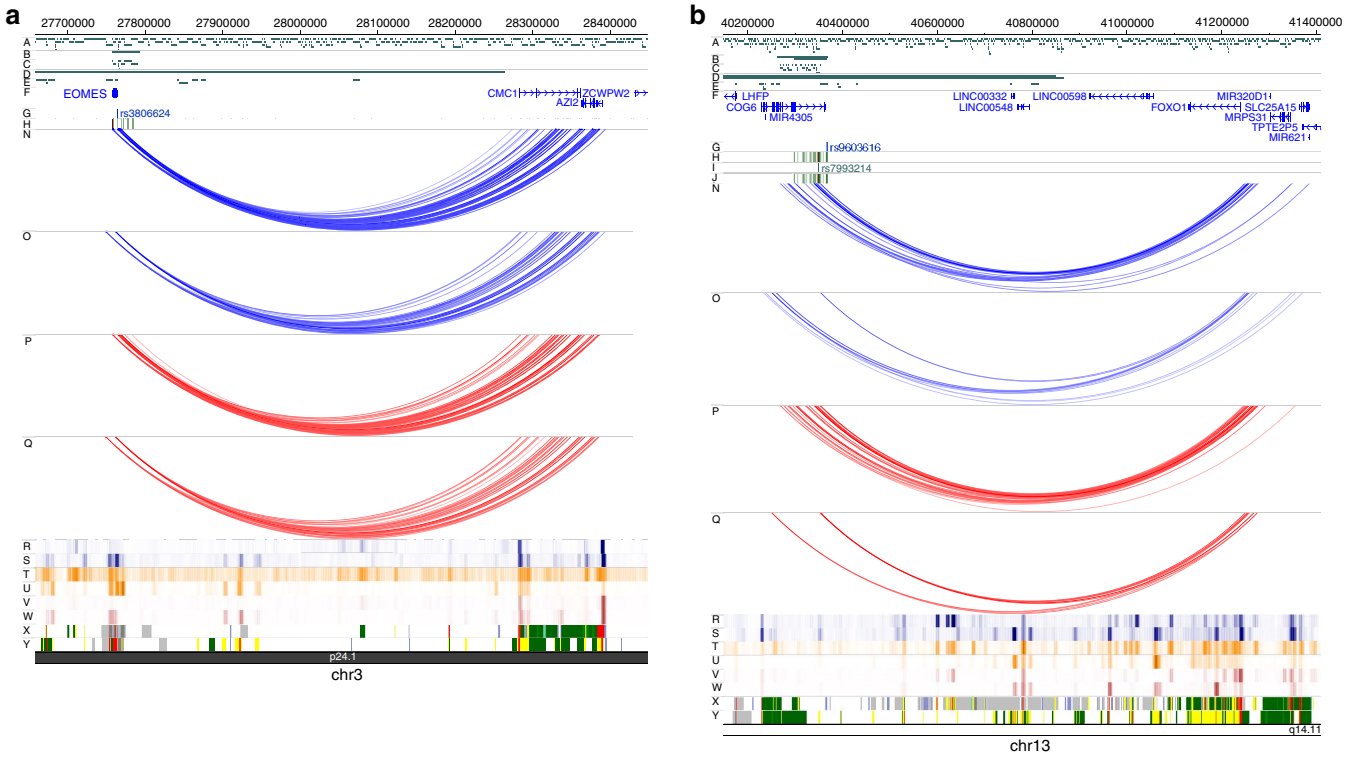


Figure 4 | Examples of chromatin interactions implicating novel gene candidates. (**a**) *EOMES* SNPs—both GM12878 and Jurkat cell lines show that SNPs situated proximal to the *EOMES* gene interact with the promoter of *AZI2I*, involved in NFκB activation, situated ~640 kb away. (**b**) *COG6* SNPs—interactions 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',

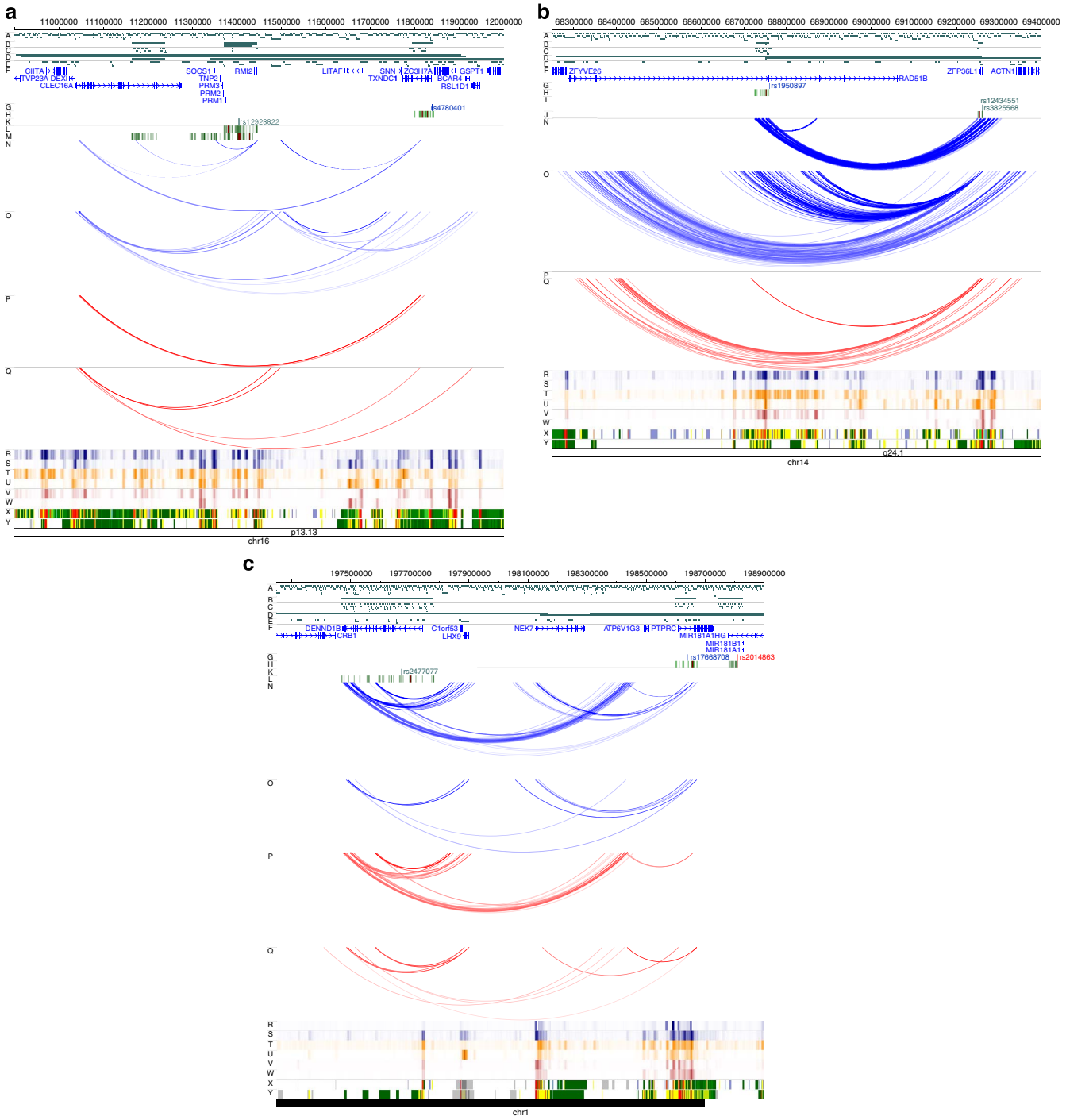


Figure 5 | Examples of chromatin interactions linking several disease associations to a common promoter. (**a**) *DEXI*—both GM12878 and Jurkat cell lines show that SNPs associated independently with RA, PsA and T1D interact with the *DEXI* promoter. In addition, evidence suggests that the RA and JIA 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 × 10^7 GM12878 and Jurkat cells were grown to ~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 × NEB Buffer 2 then re-suspended in 1.25 × NEB Buffer 2 to make aliquots of 5–6 × 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).