View metadata, citation and similar papers at core.ac.uk

brought to you by CORE

provided by Publications from Karolinska Institutet

From the DEPARTMENT OF MEDICINE, SOLNA
Karolinska Institutet, Stockholm, Sweden

GENE EXPRESSION PROFILING IN AUTOIMMUNE DISEASES – A STORY OF UPS AND DOWNS

Miranda Houtman

Karolinska Institutet

Stockholm 2018

All previously published papers were reproduced with permission from the publisher.
Cover illustration by Brian Timmer.
Published by Karolinska Institutet.
Printed by Eprint AB 2018
© Miranda Houtman, 2018
ISBN 978-91-7831-275-7

GENE EXPRESSION PROFILING IN AUTOIMMUNE DISEASES - A STORY OF UPS AND DOWNS

THESIS FOR DOCTORAL DEGREE (Ph.D.)

By

Miranda Houtman

Principal Supervisor:
Associate Professor Leonid Padyukov
Karolinska Institutet
Department of Medicine, Solna
Division of Rheumatology

Co-supervisors:
Professor Vivianne Malmström
Karolinska Institutet
Department of Medicine, Solna
Division of Rheumatology

Senior Researcher Karine Chemin
Karolinska Institutet
Department of Medicine, Solna
Division of Rheumatology

Opponent:
Associate Professor Caroline Ospelt
University Hospital Zurich
Department of Rheumatology
Center of Experimental Rheumatology

Examination Board:
Associate Professor Lars Feuk
Uppsala University
Department of Immunology, Genetics and Pathology

Associate Professor Andrey Alexeyenko
Karolinska Institutet
Department of Microbiology, Tumor and Cell Biology

Associate Professor Andor Pivarcsi
Karolinska Institutet
Department of Medicine, Solna
Division of Dermatology and Venereology

The public defense will take place on Friday the 14th of December 2018 at 9:00 am in the Center for Molecular Medicine (CMM) lecture hall, L8:00, Karolinska University Hospital, Solna.

ABSTRACT

Autoimmune diseases are believed to arise from a combination of genetic and environmental factors that affect normal function of immune cells. In this thesis, we studied the functional role of genetic variants, in peripheral blood cells, that relate to rheumatoid arthritis (RA) and myositis by gene expression profiling.

Genome wide association studies have identified numerous susceptibility loci for autoimmune diseases, however, the precise mechanisms of how these loci lead to increased risk of autoimmunity remain mostly unknown. We therefore aimed to increase our understanding of the involvement of the susceptibility loci *PTPN2*, *PTPN22* and *HLA-DRB1* in the pathogenesis of RA. For the *PTPN2* locus, we show that the long non-coding RNA (lncRNA) *LINC01882* encoded on this locus can be linked to RA. We found that the genetic variants in the *PTPN2* locus are associated with the expression of several lncRNAs, but not with the expression of *PTPN2*. By silencing *LINC01882* in Jurkat T cells, we identified that *LINC01882* might play a role in T-cell activation by regulating IL-2 levels, an important cytokine in RA. In addition, we show a new role for the *PTPN22* risk allele in the context of RA through the generation of CD4+ T cells with cytotoxic characteristics. We found that genes related to T-cell survival and cytotoxic T-cell differentiation were differentially expressed between *PTPN22* risk and non-risk allele carriers. This led us to identify an increased frequency of EOMES+CD4+ T cells in healthy individuals carrying the *PTPN22* risk allele. Furthermore, we identified a difference in the expression of *HLA-DRB1* and certain *HLA-DQ* genes between healthy individuals carrying RA *HLA-DRB1* risk (*04:01) and non-risk (*15:01) alleles. These differences in gene expression were observed in different cell types, including CD4+ and CD8+ T cells. This data suggests that *HLA-DRB* and *HLA-DQ* levels, and potentially their corresponding proteins, might support loss of immune tolerance in RA patients carrying *HLA-DRB1**04:01 alleles.

In addition, we aimed to differentiate involvement of CD4+ and CD8+ T cells in the myositis subgroups, polymyositis (PM) and dermatomyositis (DM), by studying gene expression. We found two genes that were differentially expressed in CD4+ T cells of patients with PM compared to DM, whereas we identified 176 genes that were differentially expressed in CD8+ T cells of patients with PM compared to DM. Several of these genes were related to lymphocyte migration and regulation of T-cell differentiation. These results add to the evidence that different immune mechanisms are involved in patients with PM compared to patients with DM.

In summary, this thesis presents several new mechanisms for the RA susceptibility loci *PTPN2*, *PTPN22* and *HLA-DRB1*. As these susceptibility loci are shared between several autoimmune diseases, these results can be implicated in the pathogenesis of other autoimmune diseases as well. We further suggest that different immune mechanisms are involved in subgroups of RA and myositis patients. These results could ultimately lead to the identification of more specific therapeutic targets for different autoimmune diseases.

LIST OF SCIENTIFIC PAPERS

I. T cells are influenced by a long non-coding RNA in the autoimmune associated *PTPN2* locus
Houtman M, Shchetynsky K, Chemin K, Hensvold AH, Ramsköld D, Tandre K, Eloranta ML, Rönnblom L, Uebe S, Catrina AI, Malmström V, and Padyukov L
Journal of Autoimmunity 2018;90:28-38

II. EOMES-positive CD4+ T cells are increased in *PTPN22* (1858T) risk allele carriers
Chemin K, Ramsköld D, Diaz-Gallo LM, Herrath J, Houtman M, Tandre K, Rönnblom L, Catrina A, and Malmström V
European Journal of Immunology 2018;48:655-669

III. RNA expression of *HLA-DRB* and *HLA-DQ* genes in healthy individuals differs between MHC class II haplotypes
Houtman M, Hesselberg E, Tandre K, Eloranta ML, Rönnblom L, Klareskog L, Malmström V, and Padyukov L
Manuscript

IV. T-cell transcriptomics from peripheral blood highlights differences between polymyositis and dermatomyositis patients
Houtman M, Ekholm L, Hesselberg E, Chemin K, Malmström V, Reed AM, Lundberg IE, and Padyukov L
Arthritis Research & Therapy 2018;20:188-202

ADDITIONAL PUBLICATIONS

I. An antisense RNA capable of modulating the expression of the tumor suppressor microRNA-34a
Serviss JT, Andrews N, van den Eynden J, Richter FC, Houtman M, Vesterlund M, Schwarzmueller L, Johnsson P, Larsson E, Grandér D, and Pokrovskaja Tamm K
Cell Death & Disease 2018;9:736-750

II. Reply to Liu et al.: Translation of rat congenic data to humans on a conserved MHC-III haplotype associated with rheumatoid arthritis
Yau AC, Houtman M, Padyukov L, and Holmdahl R
Proceeding of the National Academy of Sciences of the United States of America 2016;113:E6323-E6324

III. Conserved 33-kb haplotype in the MHC class III region regulates chronic arthritis
Yau AC, Tuncel J, Haag S, Norin U, Houtman M, Padyukov L, and Holmdahl R
Proceeding of the National Academy of Sciences of the United States of America 2016;113:E3716-E3724

CONTENTS

1 Introduction ... 1
1.1 Gene expression and regulation ... 1
1.1.1 Gene expression ... 1
1.1.2 RNA splicing ... 2
1.1.3 Non-coding RNAs ... 3
1.1.3.1 Small non-coding RNAs ... 3
1.1.3.2 Long non-coding RNAs ... 4
1.2 Genetic variation ... 7
1.2.1 Single nucleotide polymorphisms ... 7
1.2.2 Genome wide association studies ... 8
1.2.3 From genetic variation to function ... 8
1.2.3.1 Expression quantitative trait loci ... 9
1.2.3.2 Regulatory quantitative trait loci ... 10
1.3 Immunity ... 11
1.3.1 Innate immunity ... 11
1.3.2 Adaptive immunity ... 12
1.4 Autoimmunity ... 14
1.4.1 Rheumatoid arthritis ... 14
1.4.1.1 Risk factors ... 15
1.4.1.2 Disease mechanisms ... 17
1.4.2 Idiopathic inflammatory myopathies ... 18
1.4.2.1 Dermatomyositis ... 19
1.4.2.2 Polymyositis ... 19
1.4.3 Unmet needs in autoimmune diseases ... 19
2 Aims ... 21
3 Experimental procedures ... 23
3.1 Cohorts and methodology ... 23
3.1.1 EIRA study (Paper I) ... 23
3.1.1.1 mQTLs ... 23
3.1.2 COMBINE study (Paper I and II) ... 23
3.1.2.1 eQTLs ... 24
3.1.3 Healthy individuals – *HLA-DRB1* (Paper I and III) ... 24
3.1.3.1 Mapping to the MHC region ... 24
3.1.4 Healthy individuals – *PTPN22* (Paper I and II) ... 25
3.1.5 Myositis patients (Paper IV) ... 25
3.2 Additional methods ... 26
3.2.1 Knockdown of *LINC01882* (Paper I) ... 26
3.2.2 Quantitative real-time PCR (Paper I and II) ... 26
3.2.3 HLA typing (Paper III and IV) ... 27
4 Results and discussion ... 29

4.1 Paper I: T cells are influenced by a long non-coding RNA in the autoimmune associated *PTPN2* locus 29
4.1.1 Rationale....29
4.1.2 QTL analyses in healthy individuals and RA patients 29
4.1.3 Expression profile of the lncRNAs in T cells....30
4.1.4 Knockdown of *LINC01882* in Jurkat T cells....31
4.1.5 Concluding remarks....32
4.2 Paper II: EOMES-positive CD4+ T cells are increased in PTPN22 (1858T) risk allele carriers....32
4.2.1 Rationale....32
4.2.2 Differential expression in naïve CD4+ T cells 32
4.2.3 EOMES+CD4+ T cells in healthy individuals and RA patients 34
4.2.4 Concluding remarks....35
4.3 Paper III: RNA expression of HLA-DRB and HLA-DQ genes in healthy individuals differs between MHC class II haplotypes 35
4.3.1 Rationale....35
4.3.2 Differential expression in PBMCs 36
4.3.3 Differential expression in isolated cell subsets....37
4.3.4 Concluding remarks....38
4.4 Paper IV: T-cell transcriptomics from peripheral blood highlights differences between polymyositis and dermatomyositis....38
4.4.1 Rationale....38
4.4.2 Differential expression in CD4+ T cells 39
4.4.3 Differential expression in CD8+ T cells 40
4.4.4 Concluding remarks....41
5 Conclusions and future perspectives....43
6 Acknowledgements....47
7 References....49

LIST OF ABBREVIATIONS

ACPA	Anti-citrullinated protein antibody
ACR	American College of Rheumatology
APC	Antigen presenting cell
ARS	Aminoactyl-tRNA synthetase
aseQTL	Allele-specific expression quantitative trait locus
BCR	B-cell receptor
CAGE	Cap-analysis of gene expression
CTCF	CCCTC-binding factor
DC	Dendritic cell
DM	Dermatomyositis
dsQTL	DNase I sensitivity quantitative trait locus
eQTL	Expression quantitative trait locus
EULAR	European League Against Rheumatism
FLS	Fibroblast-like synoviocyte
GWAS	Genome wide association study
H3K27ac	Histone 3 lysine 27 acetylation
H3K27me3	Histone 3 lysine 27 trimethylation
H3K4me1	Histone 3 lysine 4 monomethylation
HLA	Human leukocyte antigen
hQTL	Histone quantitative trait locus
IFN-	Interferon-
IL-	Interleukin-
ILD	Interstitial lung disease
LD	Linkage disequilibrium
lncRNA	Long non-coding RNA
LPS	Lipopolysaccharide
MAA	Myositis associated autoantibody
MHC	Major histocompatibility complex
miRNA	MicroRNA

mQTL	Methylation quantitative trait locus
mRNA	Messenger RNA
MS	Multiple sclerosis
MSA	Myositis specific autoantibody
MTX	Methotrexate
NAT	Natural antisense transcript
ncRNA	Non-coding RNA
NK	Natural killer
NMD	Nonsense-mediated decay
PAD	Peptidylarginine deiminase
PAMP	Pathogen-associated molecular pattern
PBMC	Peripheral blood mononuclear cell
PCA	Principal component analysis
pDC	Plasmacytoid dendritic cell
piRNA	PIWI-interacting RNA
PM	Polymyositis
pQTL	Protein quantitative trait locus
RA	Rheumatoid arthritis
RISC	RNA induced silencing complex
RNA pol II	RNA polymerase II
rRNA	Ribosomal RNA
SE	Shared epitope
siRNA	Small interfering RNA
SLE	Systemic lupus erythematosus
snoRNA	Small nucleolar RNA
SNP	Single nucleotide polymorphism
snRNA	Small nuclear RNA
sQTL	Splicing quantitative trait locus
T1D	Type 1 diabetes
TCR	T-cell receptor
TF	Transcription factor

Tfh	T follicular helper
TGF	Transforming growth factor
Th	T helper
TLR	Toll-like receptor
TNF	Tumor necrosis factor
tRNA	Transfer RNA

1 INTRODUCTION

1.1 GENE EXPRESSION AND REGULATION

All cells in a multicellular organism contain the same genetic information, however, only a fraction of the genes are expressed in each cell type. This is mainly determined by the genetic information contained in the DNA. This genetic information includes both protein-coding genes and non-coding genes that can regulate the expression of protein-coding genes. Gene expression is a tightly controlled process and any alteration in gene expression might lead to the development of diseases.

1.1.1 Gene expression

Gene expression is a tightly controlled process. It involves the transcription of protein-coding genes and many non-coding genes into RNA transcripts by RNA polymerase II (RNA pol II). The RNA pol II and its associated general transcription factors (TFs) assemble at the core promoter, that typically stretches a region of approximately 80 nucleotides around the transcription start site, and initiate transcription. The three most abundant core promoter motifs are the TATA-box, initiator motif and downstream promoter element[1]. Apart from these distinct motifs, core promoters often overlap with regions with high density of CpG dinucleotides, called CpG islands. In general, methylation of CpG islands is associated with the repression of gene expression. In addition, active core promoters exhibit highly dynamic nucleosomes containing certain histone variants, which make them accessible for the transcription machinery. Furthermore, active core promoters are associated with specific epigenetic modifications of histones.

Transcription is extremely weak in the absence of regulatory elements that are distinct from the core promoter. These regulatory elements, known as enhancers, can regulate gene expression by binding TFs that recruit transcriptional cofactors. Cofactors, either activators or repressors, are often involved in post-translational modifications of the transcription machinery and surrounding nucleosomes. It is suggested that enhancers have preferences towards certain promoters[2]. Large clusters of regulatory elements form super-enhancers. These super-enhancers are enriched for the binding of cell type specific TFs and therefore play a critical role in defining cell identity[3]. In addition, super-enhancers are often found near genes that have cell type specific functions[4].

Like promoters, enhancer activity is also under the control of epigenetic modifications. Recent studies have revealed that inactive enhancers are characterized by the presence of histone 3 lysine 27 tri-methylation (H3K27me3). During activation, the H3K27me3 modification will be replaced by H3K27 acetylation (H3K27ac). In addition, poised enhancers are marked by H3K27me3 and H3K4 mono-methylation (H3K4me1)[5]. Moreover, DNA methylation is also involved in the regulation of enhancer activity. Inactive enhancers are characterized by DNA methylation, whereas active enhancers are characterized by DNA hypomethylation[6]. Interestingly, 5-hydroxymethylcytosine, the oxidized form of 5-methylcytosine, appears to correlate with active enhancers as it prevents the methylation of the cytosine[7].

Interactions between promoters and enhancers are independent of their relative distance and can even occur between different chromosomes[8]. In order to facilitate interactions between distant loci, the genome is organized in three-dimensional structures. Each chromosome occupies a discrete territory within the nucleus and each chromosome can be further organized into loop domains and compartmental domains. Loop domains depend on the activity of architectural proteins such as CCCTC-binding factor (CTCF) and cohesin, whereas compartmental domains are independent of these proteins and reflect transcriptional and chromatin states[9]. CTCF is a zinc-finger protein composed of multiple domains that can interact with DNA, RNA and proteins to control chromatin loops. These loops represent topologically structures wherein transcription takes place[10]. Next to mediating interactions between promoters and enhancers[11], CTCF can play a role in alternative splicing and recombination[12,13]. Disruption of CTCF binding sites can affect gene expression and lead to diseases[14-16].

1.1.2 RNA splicing

During transcription, the non-protein-coding introns are removed from the pre-messenger RNA (mRNA) and the protein-coding exons are joined to form the mRNA. This process is called pre-mRNA splicing and is catalyzed by the spliceosome machinery, which consists of distinct small nuclear ribonucleoproteins. The spliceosome recognizes several motifs of the pre-mRNA, such as the splice sites and branch point. The splice sites are generally represented by the dinucleotides GU at the 5’ end of the intron and AG at the 3’ end of the intron. The branch point is about 18 - 40 nucleotides upstream of the 3’ splice site. Pre-mRNA splicing is a two-step reaction. In the first step, the branch point sequence carries out a nucleophilic attack on the 5’ splice site. In the second step, the 5’ exon initiates a second nucleophilic attack on the 3’ splice site, resulting in the release of the intron as a RNA lariat and ligation of the exons[17].

For about 95% of the human genes, splicing leads to different mature mRNA products and proteins[18,19]. This process is known as alternative splicing and increases the mRNA diversity in complex organisms. Alternative splicing is controlled by the occurrence of *cis*-regulatory RNA elements, which can act as enhancers or silencers. These RNA elements can in turn recruit trans-acting proteins, such as splicing factors. Types of alternative splicing include the use of alternative 5’ and 3’ splice sites, cassette exons, retained introns and mutually exclusive exons[20].

Alternative splicing is a tightly regulated process. It can be affected by factors such as splicing factors, pre-mRNA secondary structure and chromatin organization. In addition, the elongation rate of transcription by RNA pol II, affected by the GC content of a gene, DNA methylation density, histone marks and long terminal repeats, can regulate alternative splicing[20-22]. Transcription at lower rates may increase the recognition of splice sites and leads to alternatively spliced variants[23,24]. Moreover, alternative splicing can be a cell type specific process. For example, the transmembrane phosphatase CD45, involved in regulating proximal antigen receptor-mediated signaling, undergoes extensive alternative splicing of the cassette exons 4, 5, and 6 in response to T-cell activation. Upon T-cell activation, the expression of CD45RO (skips

exon 4, 5 and 6) is increased and the expression of CD45RA (contains only exon 4) is decreased. Skipping of the three exons reduces the phosphatase activity of CD45[25].

Alternative splicing may also create isoforms that contain premature termination codons, which are targeted for degradation by the nonsense-mediated decay (NMD) pathway. The NMD pathway degrades transcripts by removing the 5' cap and subsequent 5' and 3' digestion[26]. However, it has become apparent that the NMD pathway is also important for other cellular transcripts to regulate expression levels in a post-transcriptional manner[27]. Approximately 10% of the mRNAs in mammalian cells appear to be targets for NMD[27].

1.1.3 Non-coding RNAs

Non-coding RNAs (ncRNAs) have been known for many years. The first ncRNAs, such as ribosomal RNAs (rRNAs) and transfer RNAs (tRNAs), were already described in the 1950s[28,29]. Since then, a few more ncRNAs, including small nuclear RNAs (snRNAs) involved in splicing and small nucleolar RNAs (snoRNAs) involved in the modification of rRNAs, have been discovered. In 2001, after publishing the first draft of the human genome[30,31], it became evident that only a small fraction of the human genome encodes for proteins[32]. More recently, a study revealed that 80.4% of the human genome has a biochemical function (RNA- and/or chromatin-associated)[33]. These studies have changed our view on the human genome and highlighted the importance of ncRNAs. They are involved in a wide range of processes, such as X-chromosome inactivation and epigenetic regulation[34,35]. However, the function of most identified ncRNAs remains to be characterized. Dysregulation of ncRNAs has been implicated in numerous diseases[36-40]. ncRNAs can be classified into small ncRNAs (< 200 nucleotides) or long ncRNAs (lncRNAs (> 200 nucleotides)).

1.1.3.1 Small non-coding RNAs

The most well studied group of small ncRNAs are microRNAs (miRNAs). miRNAs are 20 - 22 nucleotides in length and regulate gene expression. The first miRNA was reported in *C. elegans* in 1993[41]. The small ncRNA *lin-4* was found to be complementary to the 3' UTR of *lin-14*[41] and suppressed accumulation of the LIN-14 protein[42]. In 2000, another miRNA, named *let-7*, was reported to be important during the development of *C. elegans*[43]. Soon thereafter, it was discovered that these miRNAs were highly conserved across phylogeny from worms to humans[44,45], suggesting that miRNAs have a more general role in biology. This led to the identification of over 1900 annotated miRNAs in the human genome (miRBase). Moreover, 60% of the human protein-coding genes are targets of miRNAs[46].

miRNAs are embedded in long primary transcripts that are transcribed by RNA pol II. The majority of miRNAs are localized within introns of protein-coding and lncNRA transcripts, but some are localized within exons[47]. In addition, several miRNAs are often in close proximity to each other, but can be individually regulated at post-transcriptional level[48]. Following transcription, the pri-miRNA, containing a hairpin structure where the miRNA is embedded, is processed by the RNase III endonuclease Drosha and its cofactor DGCR8. This complex cuts the pri-miRNA into a pre-miRNA

with intact hairpin structure and a two nucleotide long 3’ overhang. The pre-miRNA is exported into the cytoplasm by exportin 5. In the cytoplasm, the RNase III endonuclease Dicer binds to the pre-miRNA, with a preference for pre-miRNAs containing a two nucleotide long 3’ overhang, and cleaves it close to the loop structure forming a small RNA duplex. This duplex is subsequently loaded onto an AGO protein to form the RNA induced silencing complex (RISC). It removes the passenger strand to form the mature RISC. The miRNA guides the mature RISC to the miRNA binding sites, which are usually located in the 3’ UTR of mRNAs. This causes either degradation (perfect match) or repression (imperfect match) of the mRNA[47]. These processes are tightly regulated and dysregulation is associated with diseases, such as cancer, neurodevelopmental disorders and autoimmune disorders[36-38].

Other small ncRNAs include small interfering RNAs (siRNAs) and PIWI-interacting RNAs (piRNAs). siRNAs are derived from exogeneous sources, such as viruses and transposons, and function in a similar way as miRNAs to degrade mRNAs. In addition, siRNAs can be used to induce short-term silencing of genes in functional experiments[49]. piRNAs are a distinct class of 23 - 30 nucleotide long RNAs that form complexes with PIWI proteins to silence retrotransposons and other genetic elements in germ line cells via mainly epigenetic mechanisms[50,51]. Although it has long been thought that piRNAs are restricted to germ line cells, recent studies have identified piRNAs in the mammalian brain[52] and human plasma[53]. More studies are needed in order to fully elucidate the role of piRNAs in humans.

1.1.3.2 Long non-coding RNAs

The majority of ncRNAs belong to the group of lncRNAs. lncRNAs are longer than 200 nucleotides and can be transcribed by RNA pol II, 5’ capped, 3’ polyadenylated and spliced. It has been shown that lncRNAs can regulate gene expression through diverse mechanisms. The first lncRNA was discovered in the late 1980s. Two imprinted genes localized to the same cluster at mouse chromosome 7, the paternally expressed *Igf2* and the maternally expressed *H19*, were identified[54,55]. *H19* was unusual as it lacks translation even though it contained small open reading frames. Although *H19* is not translated into a protein, it was found to play a crucial role in embryonic development[56]. A couple of years later, the lncRNA *Xist* was discovered in the X-inactivation center important for initiating X-chromosome inactivation. In addition, the lncRNA *Xist* can be transcriptionally repressed by the lncRNA *Tsix*, an antisense transcript that overlaps with the gene of *Xist*[57]. In the early 2000s, it was discovered that a large number of transcriptionally active regions of human chromosome 21 and 22 were not mapping to any known protein-coding gene[58,59]. After the human genome was sequenced[30,31], it became evident that most of the genome, protein-coding or not, was actually transcribed[32].

Previous mentioned findings were supported by findings from the consortia FANTOM and ENCODE. The FANTOM consortium determined the base sequences, assigned functional annotations to a set of 60,770 full-length mouse complementary DNA sequences and identified that ncRNAs are a major component of the mouse transcriptome[60]. Later, utilizing a technology named cap-analysis of gene expression (CAGE), they revealed that antisense transcription is widespread in the mammalian

genome and might contribute to the regulation of gene expression[61]. The CAGE technique sequences pol II transcribed RNAs that are capped on the 5' end to map the transcription start sites and promoters[62]. In their latest project[63], they generated an atlas of 27,919 human lncRNAs found in major human primary cells and tissues[64]. They revealed that most of the intergenic lncRNAs originate from enhancers rather than from promoters[64]. In addition, they suggest that lncRNAs are involved in multiple diseases. They found that lncRNAs that overlap with disease-associated single nucleotide polymorphisms (SNPs) are specifically expressed in cell types relevant to the disease[64] (discussed in more detail below). The FANTOM consortium now launched a new project to systematically elucidate the function of lncRNAs in the human genome.

The ENCODE consortium aimed to identify all functional elements in the human genome by utilizing high-throughput methods, such as RNA-seq, ChIP-seq and ATAC-seq. They revealed that a total of 74.7% of the human genome was transcribed into RNA[65]. In addition, the majority of lncRNAs (92%) was not translated[66]. Furthermore, they found that lncRNAs are generated through similar pathways as protein-coding genes[67]. However, expression of lncRNAs is more cell type specific. Also, lncRNAs were found to be predominantly localized to chromatin structures and the nucleus[67]. In addition, lncRNAs do not seem to be conserved, however, their promotor regions are almost as conserved as protein-coding gene promotors[67].

Although these studies have indicated the importance of lncRNAs, the functional mechanisms of the majority of lncRNAs is not clear. lncRNAs have been identified to act through diverse mechanisms. They can interact with DNA, RNA, chromatin and proteins. lncRNAs can act as scaffolds to form protein complexes[68]. lncRNAs can also act as guides to target chromatin remodeling complexes to the genome. lncRNAs can form chromosomal loops and mediate inter-chromosomal interactions. Many lncRNAs serve as precursors for small ncRNAs, in particular siRNAs. lncRNAs can also act as competing endogenous RNAs or miRNA sponges that indirectly regulate gene expression levels[69,70]. As an example, the level of the tumor suppressor gene *PTEN* can be modulated by its pseudogene *PTENP1* as they both contain binding sites for the same miRNAs[71]. Therefore, lncRNAs can influence many cellular processes, such as chromatin remodeling, transcription, splicing and translation. The following examples demonstrate the flexibility of regulatory lncRNAs.

Chromatin remodeling

lncRNAs are involved in epigenetic modifications, including histone acetylation, histone methylation and DNA methylation. For example, the before mentioned lncRNA *Xist* coats the X-chromosome to recruit chromatin remodeling complexes to induce silencing. In more detail, *Xist* recruits the polycomb repressive complex 2 (PRC2), including the histone methyl transferase EZH2, which catalyzes formation of the repressive histone mark H3K27me3 and silences the X-linked genes (Figure 1)[72]. Moreover, the antisense transcript *Tsix* suppresses the expression of *Xist* through the induction of DNA methylation. *Tsix* interacts with the DNA methyl transferase DNMT3a and is recruited to the promoter of *Xist* to methylate it[73]. Another example is the lncRNA *HOTAIR*, encoded on the chromosome 12 *HOXC* locus. The 5' end of *HOTAIR* recruits the PRC2 complex to the *HOXD* locus on chromosome 2 to silence it

by catalyzing the formation of H3K27me3. It has been suggested that *HOTAIR* recruits PRC2 to induce silencing of multiple genes[68]. In addition, the 3' end of *HOTAIR* interacts with the CoREST/REST complex, including the histone demethylase LSD1, which demethylates H3K4me2 (a marker of actively transcribed regions). Thus, *HOTAIR* serves as a scaffold to assemble histone modifiers for silencing of *HOX* genes[68].

Transcription

lncRNAs can regulate gene transcription in several ways. One of these ways is acting as a co-regulator to enhance or repress gene expression. For example, the lncRNA *Evf2*, transcribed as a spliced anti-sense transcript from the *Dlx5/6* enhancer, forms a complex with the TF DLX2 and MECP2 and regulates the activity of the *Dlx5/6* enhancer by inhibiting DNA methylation levels (Figure 1)[74,75]. In addition, many lncRNAs are transcribed from active enhancers. These enhancer-associated lncRNAs have been found to be associated with enhanced levels of their neighboring protein-coding genes[76]. Another way is by forming a RNA-DNA triple helix to repress gene expression. For example, the *DHFR* gene contains two promoters, a general promoter and a minor promoter upstream of *DHFR*. The minor promoter transcribes a lncRNA that forms a RNA-DNA triple complex with the sequences of the major promoter and interferes with TFIIB to repress the expression of *DHFR*[77].

Splicing

lncRNAs can also affect splicing, including alternative splicing. For example, the lncRNA *MALAT1* regulates the concentration of splicing factors in nuclear speckle domains. It has been shown that *MALAT1* can affect the concentration, distribution and activity of the serine/arginine splicing factor (Figure 1)[78]. Depletion of *MALAT1* changes the alternative splicing profile of multiple pre-mRNAs. Another example is the *Zeb2* natural antisense transcript (NAT), which is upregulated after Snail1-induced epithelial mesenchymal transition. *Zeb2* NAT is complementary to the 5' splice site of an intron of the *Zeb2* mRNA, a transcriptional repressor of E-cadherin. By masking this splice site, the translation machinery recognizes an internal ribosome entry site in the intron and retains it. This results in higher levels of the Zeb2 protein and consequently lower levels of E-cadherin mRNA and protein[79].

Translation

lncRNAs can influence translation by acting directly on their target mRNAs and affect their stability. For example, the antisense transcript of *BACE1* (*BACE1-AS*) is important for the stability of *BACE1* mRNA, the precursor of an enzyme that processes the amyloid precursor protein into β-amyloid peptides involved in Alzheimer's disease. The lncRNA *BACE1-AS* can form perfect base pairs with *BACE1* mRNA to increase its stability. Moreover, the overlap between *BACE1-AS* and *BACE1* includes a target site for miR-485-5p. This miRNA can also bind to *BACE1* mRNA, but decreases the stability of *BACE1* mRNA. Thus perfect base pairing of *BACE1-AS* with *BACE1* protects the mRNA from degradation by the miRNA (Figure 1)[80]. lncRNAs can also affect the NMD pathway. *1/2-sbsRNAs* can bind to the 3'UTR of mRNA containing Alu elements. Imperfect base pairing forms binding sites for the RNA binding protein

STAU1[81]. STAU1 can cause mRNA decay by binding to the NMD factor UPF1 and bringing it to the 3'UTR of the mRNA[82]. For example, one of the lncRNAs, *1/2-sbsRNA1*, binds to the 3'UTR of *SERPINE1* and reduces the mRNA levels[81].

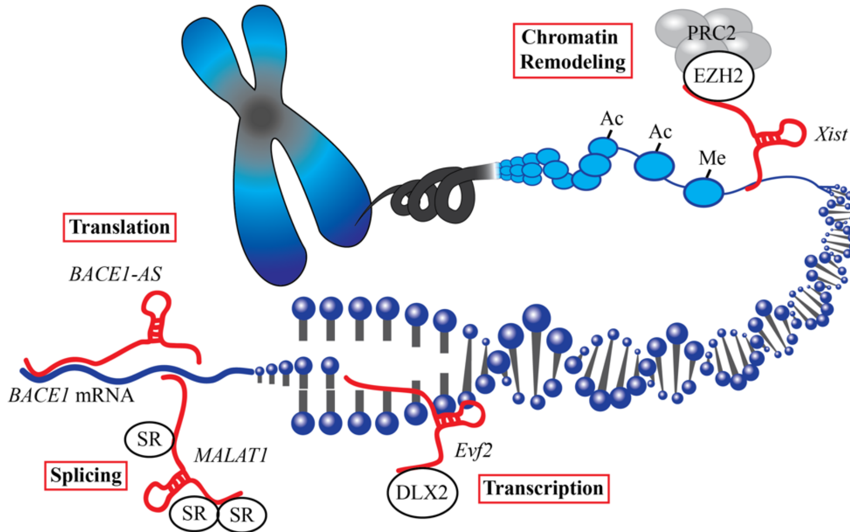

Figure 1. Models of lncRNA functions. lncRNAs can recruit chromatin remodeling complexes (Xist), regulate transcription (Evf2), influence pre-mRNA splicing (MALAT1) and regulate mRNA stability (BACE1-AS). SR: splicing regulator.

1.2 GENETIC VARIATION

New advances in sequencing and genotyping technologies have allowed scientists to look across the genome. These developments have identified millions of genetic variants in the human genome. Some of these genetic variants have been identified to be associated with human diseases and traits through genome wide association studies (GWAS). The challenge now is to unravel the mechanisms behind these associations.

1.2.1 Single nucleotide polymorphisms

Decoding the DNA sequence was one of the major scientific challenges of the last decades. The Human Genome Project published the first draft of the human genome in 2001[30]. However, genetic variation was not annotated in this version. Subsequent large-scale projects, such as the 1000 Genomes Project[83,84] and the International HapMap Consortium[85,86], were initiated to identify and catalogue genetic variation. The most common genetic variation in humans are SNPs. A SNP is defined as a single nucleotide change that is present in at least 1% of the general population. In 2015, over 84 million validated SNPs have been identified in humans[87]. This number has now passed the 110 million (dbSNP build 151). SNPs are distributed throughout the genome. SNPs located in protein-coding regions can be synonymous (do not alter the encoded amino acid sequence) or non-synonymous (do alter the encoded amino acid sequence) and can

induce changes or loss of protein function[88,89]. SNPs located in non-coding regions have the potential to be regulatory and can influence for example pre-mRNA splicing, TF binding and DNA methylation (discussed in more detail below)[90-92].

1.2.2 Genome wide association studies

Using high-throughput genotyping technologies, GWAS have identified SNPs that are associated with many complex diseases. These studies look for SNPs with significant allele frequency differences between cases and controls. GWAS rely on the correlation structure of the genome, known as linkage disequilibrium (LD). Therefore, GWAS arrays contain only a few hundred thousand tag SNPs that represent SNPs in the same LD block.

The first GWAS was published in 2002[93]. A candidate locus on chromosome 6p21, including the lymphotoxin-α gene, was identified to be associated with myocardial infarction in a Japanese population. Many GWAS investigating different diseases followed. In 2005, a GWAS was published where SNPs in the *CFH* gene (encoding complement factor H) were identified to be associated with age-related macular degeneration[94]. DNA resequencing of this gene revealed a common coding variant that increases the risk for age-related macular degeneration. The first large GWAS, including 14,000 cases of seven common diseases (coronary heart disease, type 1 diabetes (T1D), type 2 diabetes, rheumatoid arthritis (RA), Crohn's disease, bipolar disorder, and hypertension) and 3,000 shared controls, was published by the Wellcome Trust Case Control Consortium in 2007[95]. This study identified 24 independent association signals underlying these diseases. Since these studies, over 70,000 associations have been published across many diseases and traits (NHGRI-EBI GWAS Catalog[96,97]).

The primary goal of GWAS is to better understand complex diseases, however, the path from GWAS to biology is not straightforward. First, the effect of associated SNPs on disease risk is small. Second, the associated SNP is not necessarily the true causal SNP, it is likely to be in LD with the causal SNP. Fine mapping could pinpoint to the causal SNP. To fine map established GWAS loci, custom genotyping chips, such as the Immunochip[98] (designed for studying immune related diseases) and Metabochip[99] (designed for studying metabolic, cardiovascular and anthropometric traits), have been developed. For example, the Immunochip contains around 195,000 SNPs from 186 distinct associated loci from 12 immune related diseases. Third, the associated SNP does not tell which gene is implicated. Recent studies have shown that the gene can be located far from the associated SNP (even on different chromosomes)[100,101]. Fourth, the vast majority of disease-associated SNPs are located in non-coding regions and have the potential to be regulatory[102-104].

1.2.3 From genetic variation to function

GWAS have identified thousands of disease-associated SNPs, but relatively little is known about the underlying mechanisms. Several consortia, including ENCODE[33], FANTOM[105], the NIH Roadmap Epigenomics Mapping Consortium[106] and the Genotype-Tissue Expression (GTEx) project[107], collected large amounts of data to help

improve the functional annotation of regulatory SNPs. This data includes information about gene expression, pre-mRNA splicing, protein expression, chromosomal conformation, chromatin accessibility, histone modifications, DNA methylation and TF binding.

1.2.3.1 Expression quantitative trait loci

Gene expression levels can be quantified easily for thousands of genes at once, for example by conducting RNA-seq. Genetic variants that affect gene expression levels are identified as expression quantitative trait loci (eQTLs). Several studies have shown an overrepresentation of eQTLs among GWAS loci[108-110], suggesting that disease-associated variants impact cellular phenotypes. Most of the observed eQTLs are located near the affected gene (usually within 1 megabase), known as local eQTLs (*cis*-eQTLs). *Cis*-eQTLs show widespread sharing across populations[111] and are often located in regulatory elements, such as promoters and enhancers. In addition, lncRNA loci are enriched for *cis*-eQTLs[110].

There have been several studies that show that *cis*-eQTLs can aid in pinpointing the causal variant within a locus. For example, common SNPs in the *FTO* locus have been found to be associated with obesity in different populations[112-114]. Multiple studies have focused on unraveling potential mechanisms, however, mostly unsuccessful as there were no SNPs that disrupted the protein-coding sequence of FTO. A recent study discovered that the SNP rs1421085, located in the first intron of *FTO*, affects the expression of *IRX3* and *IRX5*[115]. Functional studies suggested that *IRX3* plays a role in obesity. However, for many other disease-associated variants the potential causal gene is still unknown as most of the common genetic variants are actually associated with one or more genes[110,116]. This implies that identifying the causal gene requires further investigation. In addition, the affected gene may be located several megabases away from the genetic variant, known as distal eQTLs (*trans*-eQTLs).

The number of reported *trans*-eQTLs is much lower than the number of reported *cis*-eQTLs. This is most likely a consequence of the small sample size of the initial eQTL mapping studies. However, *trans*-eQTL analyses can provide valuable insight into disease pathogenesis[117-119]. For example, a SNP in the *IRF7* locus, associated with the autoimmune disease T1D, affects the expression of *IRF7* in *cis*[117]. After viral exposure, it affects the expression of seven other genes in *trans*. This suggested a role of the IRF7-driven inflammatory network in the etiology of T1D. A second example is the SNP rs4917014 in the *IKZF1* locus, which is associated with the autoimmune disease systemic lupus erythematosus (SLE). This SNP affects the expression of *IKZF1* in *cis* and the expression of five type I interferon-α (IFN-α) response genes and four genes involved in complement in *trans*[119]. These genes have increased binding of the TF encoded by *IKZF1*, suggesting the importance of *IKZF1* in SLE.

Genetic variants act via molecular pathways to alter phenotypes, which suggest that the effect of these variants will be specific to certain cell/tissue types relevant to the phenotype. Indeed, many studies have found that eQTLs can be cell type specific[110,120-122]. This is especially true for *trans*-eQTLs, while 25.3% of *cis*-eQTLs were shared

across three or more tissues, only 3.8% of the *trans*-eQTLs were shared[110]. The *trans*-eQTLs were mostly shared across a subset of related tissues, such as different brain regions[110]. However, eQTLs can be specific even across closely related cell types. In the above described obesity example, the SNP rs1421085 was found to affect the expression of *IRX3* in pre-adipocytes but not in whole adipose tissue[115]. Therefore, this field is moving from heterogenous cell type samples, such as whole blood, to specific cell types, such as monocytes[123] and regulatory T cells[124].

In addition to cell type specificity, eQTLs can be context specific. Some eQTLs might only be detectable under specific conditions, such as environmental stimuli. For example, a study in dendritic cells, assessing the effect of bacterial lipopolysaccharides (LPS), influenza and interferon-β (IFN-β) on gene expression, reported that about half of the observed eQTLs were only discovered in stimulated cells[125]. Similarly, a study in monocytes, mapping eQTLs before and after stimulation with bacterial LPS and interferon-γ (IFN-γ), detected that more than half of the *cis*-eQTLs were context specific[123]. They found that *IFNB1*, encoding the cytokine interferon B1, is among the *cis*-eQTLs in monocytes that is induced after 2 hours of LPS stimulation. However, after 24h of LPS stimulation, this *cis*-eQTL is no longer observed, but in turn multiple *trans*-eQTLs appear that are downstream targets of IFNB1.

In addition to total gene expression (gene expression as an average over all alleles), allele-specific expression can be used to study eQTLs. Allele specific expression quantitative trait loci (aseQTLs) provide strong evidence of *cis*-regulatory mechanisms, whereas *cis*-eQTLs can arise from non-*cis* mechanisms[126].

Genetic variants can also be associated with alternative splicing, referred to as splicing quantitative trait loci (sQTLs). sQTLs tend to be enriched at 5' and 3' splice sites, but many SNPs located outside of these splice site regions have been associated with alternative splicing as well. These SNPs can for example modify regulatory RNA elements, such enhancers and silencers. Several studies have reported an enrichment of sQTLs among disease-associated variants[127,128], suggesting that variants affecting alternative splicing could be the causal variants underlying GWAS signals. As an example, the SNP rs2248374, located close to a 5' splice site of *ERAP2*, is associated with several diseases, including Crohn's disease[129]. It deactivates the canonical 5' splice site of *ERAP2* and in turn activates a downstream splice site. This introduces a premature stop codon and the transcript is targeted by the NMD pathway for degradation[90]. Therefore, alternative splicing might be underlying some of the eQTLs.

Besides eQTLs, protein quantitative trait loci (pQTLs) can be used to pinpoint potential causal variants from a GWAS[130,131]. pQTLs are often driven by regulation of mRNA levels, but some pQTLs may be involved in post-translational regulation[132,133].

1.2.3.2 Regulatory quantitative trait loci

Epigenomic mechanisms, such as chromatin accessibility, histone modifications, DNA methylation and TF binding, can also underlie eQTLs. Genetic variants associated with chromatin accessibility, studied by mapping DNase I hypersensitivity sites, are called

DNase I sensitivity quantitative trait loci (dsQTLs). A study in lymphoblastoid cell lines found that 55% of the eQTLs are also dsQTLs[134]. In addition, they identified that dsQTLs are enriched within TF binding sites. A further study used histone modifications to detect differences in chromatin state associated with genetic variants, resulting in histone quantitative trait loci (hQTLs)[135]. Moreover, multiple studies have identified genetic variants that are associated with DNA methylation, known as methylation quantitative trait loci (mQTLs)[136-138]. Most of the mQTLs act in *cis* and mQTLs are enriched in TF and CTCF binding sites[138]. Taken together, dsQTLs, hQTLs and mQTLs mainly point to the disruption of TF binding sites as the mechanism driving the associations.

In the above described obesity example, the SNP rs1421085, located in the first intron of *FTO*, was found to affect the expression of *IRX3* and *IRX5*. The risk allele disrupts a motif for the repressor ARID5B, which leads to the derepression of an enhancer and increased expression of *IRX3* and *IRX5*[115]. Similarly, the SNP rs968567, located within the promoter of *FADS2*, is associated with RA. The risk allele disrupts the binding of the TF SREBF2, which down-regulates the expression of *FADS2*[91]. It is suggested that methylation of nearby CpG sites affect the binding of SREBF2[139]. These examples illustrate that integrating different molecular traits can uncover the pathways underlying GWAS hits.

1.3 IMMUNITY

Multicellular organisms have a sophisticated defense system to prevent and fight infections without harming the host. This system consists of two mechanisms: innate immunity, which is the initial mechanism against infections, and adaptive immunity, which is a more specialized and effective mechanism against infections.

1.3.1 Innate immunity

The innate immune system has several types of defensive barriers. The first barrier is the epithelial layer of the skin, gut and lungs, which physically interfere with the entry of pathogens. Pathogens do occasionally breach the epithelial barrier. When pathogens invade the tissue, it is up to the immune system to recognize and destroy them. The innate immune system recognizes particular structures that are shared by many pathogens but are absent in the host, known as pathogen-associated molecular patterns (PAMPs)[140]. The PAMPs are recognized by membrane-bound and cytoplasmic receptors in the host, such as Toll-like receptors (TLRs), C-type lectin receptors and nucleotide-binding oligomerization domain-like receptors[141]. They can also be recognized by circulating receptors, such as components of the complement system[142].

The tissue resident innate immune cells, such as macrophages, dendritic cells (DC) and in some cases mast cells, are among the first cells to encounter invading pathogens. These cells can get activated through the above-mentioned receptors and phagocyte the pathogen. This leads to the production of a variety of inflammatory mediators, such as cytokines (e.g. tumor necrosis factor α (TNF-α) and interleukin-1 (IL-1)) and chemokines (e.g. C-C motif chemokine ligand 2 (CCL2)). These inflammatory

mediators increase the permeability of the blood vessel, which facilitates the passage of recruited neutrophils and other leukocytes into the tissue to promote more inflammation[140].

Defense against viruses include special mechanisms that involve type I interferons (IFN-α and IFN-β) and natural killer (NK) cells. Although type I interferons can be produced by almost any cell type in the body, plasmacytoid DCs (pDCs) are specialized for the production of a large amount of type I interferons[143]. In addition, type I interferons are secreted by virus infected cells to induce cell-intrinsic antiviral states in infected cells and neighboring uninfected cells, which limit the spread of viral pathogens. Infected cells reduce the expression of major histocompatibility complex (MHC) class I on their cell surface, which stimulate NK cells to eliminate the infected cells by releasing cytotoxic granzymes and perforins[144]. Furthermore, type I interferons activate the adaptive immune system.

1.3.2 Adaptive immunity

The innate immune system stimulates the adaptive immune system to recruit more specialized components to eliminate the pathogen. Macrophages and DCs (and B cells), known as antigen presenting cells (APCs), display antigens (peptides) on MHC molecules to T cells. T cells recognize these complexes by their expressed T-cell receptor (TCR) and proliferate rapidly to produce a large number of cells with a certain specificity. B cells can recognize with their expressed B-cell receptor (BCR) a wide variety of antigens in their native form, including proteins, lipids and toxins, and can differentiate in response to an antigen into cells that secrete antibodies[140].

T-cell precursors are derived from hematopoietic stem cells in the bone marrow and migrate to the thymus for maturation into naïve CD4+ and CD8+ T cells. During T-cell maturation, TCRs undergo genetic recombination (V(D)J recombination) to increase their diversity. Each T cell expresses a single type of TCR that can bind to a specific peptide. Thereafter, naïve CD4+ and CD8+ T cells circulate to secondary lymphoid organs where they encounter APCs expressing MHC-peptide complexes. Naïve CD4+ T cells recognize MHC class II molecules present on APCs that display extracellular peptides, whereas naïve CD8+ T cells recognize MHC class I molecules present on all nucleated cells that display intracellular peptides[140]. However, cross-presentation has been demonstrated as well[145]. In addition, costimulatory molecules (CD80/CD86 on APCs and CD28 on T cells) are needed to fully activate the T cells[146] (Figure 2). This results in clonal expansion of the T cells and differentiation into different T-cell subsets defined by the expression of specific master TFs and signature cytokines. Depending on the cytokine environment, CD4+ T cells can differentiate into T helper 1 (Th1) cells, Th2 cells, Th17 cells, T follicular helper (Tfh) cells and regulatory T cells. Th1 cells are primarily induced by IFN-γ and IL-12 and secrete IFN-γ promoted by the TF T-bet[147]. In addition, IL-2 and TNF-α are produced as well. Th1 cells are crucial in the defense against intracellular viral and bacterial pathogens. Th2 cells are dependent on IL-2 and IL-4 for their differentiation[147]. The TF GATA3 promotes the expression of IL-4, IL-5 and IL-13, whereas it suppresses the expression of IFN-γ[148]. Th2 cells play an important role in immune responses against large extracellular pathogens, such as

helminths. Th17 cells are developed from naïve CD4+ T cells in the presence of IL-6 and IL-1β and are expanded in the presence of IL-23 and transforming growth factor beta (TGF-β). The TF RORγt promotes the expression of IL-17 and IL-22 that are secreted by Th17 cells[147]. These cells are important for immune responses towards specific fungi and extracellular bacteria. In addition to these Th cell subsets, Tfh cells are specialized providers of T cell help to B cells in the germinal center. Tfh cells are induced by IL-6 and IL-21, which promote the expression of the TF BCL6. In turn, BCL6 suppresses the expression of factors that mediate the differentiation of Th1, Th2 and Th17 cells. Tfh cells primarily secrete IL-21 and IL-4[149]. Moreover, naïve CD4+ T cells can differentiate into regulatory T cells depending on TGF-β and IL-2. The master TF of regulatory T cells is FOXP3. Regulatory T cells have a crucial role in the maintenance of immunological tolerance to self and foreign antigens by for example secreting anti-inflammatory cytokines, such as IL-10 and TGF-β[150]. CD8+ T cells differentiate into cytotoxic T cells, which are involved in the destruction of infected cells expressing the antigen by the release of granzymes and perforins[151]. Some of the CD4+ and CD8+ develop into memory T cells, which will respond rapidly to repeated exposure of the same pathogen[140].

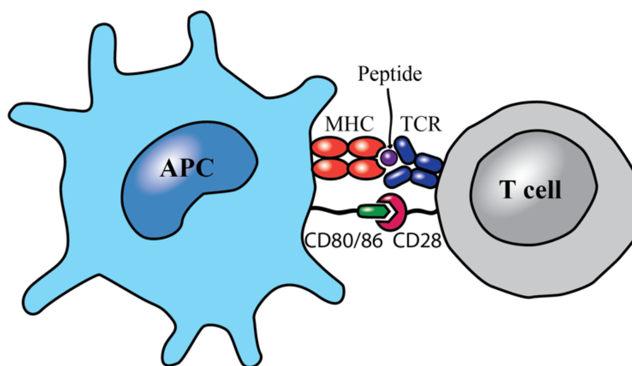

Figure 2. Interaction between an APC and T cell. Naïve CD4+ T cells recognize MHC class II molecules on APCs, whereas naïve CD8+ T cells recognize MHC class I molecules on APCs.

B cells are produced in the bone marrow and are further matured in secondary lymphoid organs. Like TCRs, BCRs undergo V(D)J recombination to increase their diversity. B cells can recognize antigens directly (without MHC molecules) and can get activated under the influence of Th cells or other stimuli. This leads to clonal expansion of the B cell and differentiation into antibody-secreting plasma cells. BCRs of activated B cells can class switch its constant region to alter its effector function. In addition, BCRs of activated B cells can undergo a process known as somatic hypermutation to improve their ability to bind an antigen. Some activated B cells become long-lived B memory cells. Plasma cell secreted antibodies are involved in neutralizing pathogens and toxins, opsonizing and phagocytosis of pathogens and antibody-dependent cellular cytotoxicity[140].

1.4 AUTOIMMUNITY

One of the most important features of the adaptive immune system is to remain unresponsive to self-antigens, known as immune tolerance. It can be differentiated into central and peripheral tolerance. Central tolerance deletes, by apoptosis or anergy, high affinity autoreactive T cells before they develop into fully matured T cells[152]. Importantly, the TF AIRE controls the expression of self-antigens in the thymus that are usually only expressed in peripheral tissues[153]. However, some autoreactive T cells can escape thymus selection and mechanisms of peripheral tolerance will prevent their activation. For instance, regulatory T cells can inhibit the activation of potentially harmful naïve T cells and their differentiation into effector T cells[154].

The complex network that maintains immune tolerance can experience defects that lead to autoimmune diseases. Autoimmune diseases are characterized by unprovoked activation of the adaptive immune system, as well as the development of autoreactive T cells and/or autoantibodies, which leads to tissue injury. It is suggested that autoimmunity is a consequence of genetic and environmental factors and failed immune regulation[155]. Examples of autoimmune diseases are RA and idiopathic inflammatory myopathies.

1.4.1 Rheumatoid arthritis

RA is a chronic inflammatory disease that is characterized by inflammation and major damage of the synovium, which leads to joint swelling, bone destruction, pain and long-term disability. The inflammation of the synovium is characterized by infiltrating cells dominated by CD4+ T cells and macrophages[156,157]. In addition, matrix-metalloproteinases and pro-inflammatory cytokines contribute to the process of cartilage and bone destruction[157]. Furthermore, RA is characterized by circulating autoantibodies. The rheumatoid factor, an autoantibody to the Fc fragment of immunoglobulin G molecules, was first observed[158]. More recently, anti-citrullinated protein antibodies (ACPAs), antibodies directed against peptides and proteins that are citrullinated, were described in the serum of up to 70% of the RA patients[159]. Citrullination is a process where the amino acid arginine within the peptide sequence is replaced with citrulline by the enzyme peptidylarginine deiminase (PAD). Although the joint is the main target for inflammatory processes in RA, there are a number of extra-articular manifestations, including cardiovascular disease and interstitial lung disease (ILD)[160].

RA affects around 0.5 - 1% of the population in western countries and the incidence of RA increases with age. Like most other autoimmune diseases, women are more often affected than men (2.5-3.0/1)[161]. There is no cure for RA. Current treatment regimens aim at achieving clinical remission at an early stage of disease[162]. Therefore, it is crucial to diagnose patients early. To facilitate this, the American College of Rheumatology (ACR) and the European League Against Rheumatism (EULAR) designed classification criteria based on joint involvement, serology, acute-phase reactants and duration of symptoms[163].

1.4.1.1 Risk factors

The pathogenesis of RA is complex and implies several risk factors. The genetic component of RA was discovered from a number of twin studies. These studies show a concordance rate for RA of 12 - 15% for monozygotic twins compared to 2 - 5% for dizygotic twins and around 1% for the general population[164-166]. However, these studies implicate that many other factors, including environmental factors, are involved in the pathogenesis of RA. Epigenetics might have a role in the pathogenesis of RA as well. In addition, interactions between these factors can have a more significant impact on the overall risk of RA.

Genetics

Multiple specific genetic loci have been identified that are associated with RA. As for many other autoimmune diseases, the strongest genetic risk factor for RA is found in the human leukocyte antigen (HLA) locus, in particular the *HLA-DRB1* gene. Over 100 other genetic factors have also been associated with RA[167]. However, most of these loci have low effect sizes, with odds ratios typically less than 1.3, and our understanding of their role in the pathogenesis of RA is limited. One of the strongest associations, next to *HLA-DRB1*, is with SNPs in the *PTPN22* locus. One of the recently found associations exist with SNPs in the *PTPN2* locus[168,169].

HLA-DRB1 – The MHC class II gene *HLA-DRB1* is the strongest genetic risk factor for RA and many other autoimmune diseases. *HLA-DRB1* encodes the beta chain of the HLA-DR heterodimer (the alpha chain of the heterodimer is encoded by the MHC class II gene *HLA-DRA*). In addition to *HLA-DR* genes, the MHC class II locus on chromosome 6 is comprised of *HLA-DP* and *HLA-DQ* genes, encoding alpha and beta chains, which are also expressed as heterodimers on APCs. As mentioned above, these heterodimers on APCs present antigens (foreign and self) to CD4+ T cells, which become activated and stimulate B cells to produce antibodies. Multiple alleles of the *HLA-DRB1* gene (*HLA-DRB1**01, *04 and *10) have been found to be associated with RA, especially ACPA-positive RA. These so-called shared epitope (SE) alleles share sequences encoding five amino acids in position 70 - 74 of the third hypervariable region of the HLA-DR beta chain[170]. However, the SE region faces mainly away from the antigen-binding groove and the role of *HLA-DRB1* SE alleles in antigen presentation was doubted. A more recent study found that amino acid 11 (or 13, as it is tightly linked to amino acid 11) of the antigen-binding groove of the HLA-DR beta chain was associated with ACPA-positive RA[171]. It has now been suggested that citrullinated peptides, bind to the positively charged peptide binding groove of the HLA-DR beta chain (mainly *HLA-DRB1**04:01 and *04:04 alleles) and will be presented to CD4+ T cells[172].

The *HLA-DRB1* SE alleles are not only related to RA susceptibility, they have been found to be associated with disease severity as well[173-175]. Patients with *HLA-DRB1* SE alleles are more likely to develop joint damage and severe forms of RA compared to patients without *HLA-DRB1* SE alleles. In addition, patients with double *HLA-DRB1* SE alleles are at an even higher risk of developing severe forms of RA[175].

HLA-DRB1 SE alleles are associated with increased risk of ACPA-positive RA. A meta-analysis discovered that *HLA-DRB1**13:01 alleles provide protection against ACPA-positive RA[176]. However, *HLA-DRB1**13 alleles in combination with the *HLA-DRB1**03 alleles increase the risk of ACPA-negative RA[177].

Several interactions between *HLA-DRB1* SE alleles and other RA risk loci have been identified, including *PTPN22*[178,179] and *MAP2K4*[180]. For example, individuals carrying the *HLA-DRB1* and *PTPN22* risk allele were more susceptible to RA compared to individuals carrying one of these risk alleles[178]. Recently, an enrichment of interactions between *HLA-DRB1* SE alleles and non-HLA risk SNPs was identified in ACPA-positive RA[179].

PTPN22 – The gene *PTPN22*, encoding the phosphatase LYP, is a strong susceptibility gene that is shared by many autoimmune diseases, including T1D[181,182], SLE[182,183], polymyositis[184] and RA[182,185]. The *PTPN22* risk allele (rs2476601T), an amino acid substitution of arginine (R) to tryptophan (W) in codon 620 (R620W) of the non-catalytic part of LYP, doubles the risk of developing ACPA-positive RA but not ACPA-negative RA[186]. Interestingly, the *PTPN22* risk allele is not associated with multiple sclerosis (MS)[182,187] and is protective in Crohn's disease[188]. LYP is primarily expressed in lymphoid cells. In T cells, LYP has been identified as a negative regulator of TCR signaling by interacting with c-Cbl, CSK, VAV, Lck, Zap70, TCR zeta chain and Grb2[189-192]. However, the functional impact of LYP620W is still incompletely understood, as studies in T cells have described that LYP620W could be a gain-of-function and a loss-of-function variant[193,194]. Moreover, some findings derived from mouse studies were not in line with the findings derived from human studies. For example, two studies showed that knock-in mice expressing the LYP variant homolog PEP619W had an increased TCR signaling and an expansion of effector-memory T cells[195,196]. There are also studies suggesting that LYP620W could influence the innate immune response by modifying TLR signaling[197].

PTPN2 – One of the recently identified candidate genes for RA is *PTPN2*. Several non-coding SNPs in the *PTPN2* locus have been reported to be associated with RA, juvenile idiopathic arthritis, inflammatory bowel disease and celiac disease[95,198-203]. The gene *PTPN2* encodes a phosphatase from the same family as LYP. The expression of *PTPN2* was first identified in T cells, but *PTPN2* is expressed in other lymphoid cells as well. PTPN2 is suggested to play a critical role in controlling immune reactions, as *Ptpn2*-deficient mice die a few weeks after birth of systemic inflammation marked by excessive production of cytokines, including TNF and IFN-γ, and nitric oxide[204]. However, the cell types that contribute to this phenotype remain widely unknown. PTPN2 has been found as an important negative regulator of cytokine signaling, including IL-2 and IFN-γ signaling, by dephosphorylating JAK1, JAK3, STAT1, STAT3 and STAT5 molecules[205-208]. In addition, PTPN2 negatively regulates TCR signaling by dephosphorylating the Src family kinases Lck and Fyn[209]. Although many studies in mice and human have shown the importance of PTPN2 in autoimmunity, the exact mechanisms by which autoimmune disease-associated SNPs in the *PTPN2* locus influence PTPN2 are not clear.

Environment

Multiple environmental factors have been associated with RA. The most clearly demonstrated environmental risk factor is smoking. Tobacco smoking increases the risk of RA and the effect is dose dependent[210-213]. The association with smoking is only found in ACPA-positive RA[211]. In addition, a gene-environment interaction between smoking and *HLA-DRB1* SE alleles has been reported for ACPA-positive RA[214]. Another study identified that smoking interacts with several non-HLA risk loci, including *PTPN22* and *PADI4*[215]. In addition to smoking, other respiratory factors, such as exposure to silica dust, mineral oil and textile dust, are associated with an increased risk of developing ACPA-positive RA[216-219]. The association with textile dust was also observed for ACPA-negative RA[219]. Obesity has been associated with an increased risk of developing RA[220-222], whereas alcohol consumption was found to be modestly associated with an reduced risk of developing RA[223,224]. Furthermore, periodontal disease is associated with an increased risk of developing RA[225,226]. The microbiota of the gut and lung might also play a role in the development of RA[227,228]. Virus infections such as Epstein-Barr virus have been associated with RA as well[229].

Epigenetics

Epigenetic mechanisms, including DNA methylation and histone modifications, might be implicated in RA. A large number of loci have been identified to be differentially methylated in patients with RA compared to osteoarthritis patients or healthy individuals (mainly by studying fibroblast-like synoviocytes (FLSs))[230-233]. Interestingly, different DNA methylation patterns have been observed in FLSs from different joints of patients with RA[234]. In addition, DNA methylation might mediate gene-environment interactions in a cell type specific manner. For example, smokers with ACPA-positive RA carrying *HLA-DRB1* SE alleles have higher DNA methylation levels than smokers with ACPA-positive RA who do not carry *HLA-DRB1* SE alleles[235]. The difference in methylation was not observed in nonsmokers[235]. Furthermore, inhibitors of histone deacetylase have been described to have beneficial effects in mouse models of arthritis[236,237], suggesting the importance of histone acetylation in RA.

1.4.1.2 Disease mechanisms

The initiation of RA is probably years before clinical symptoms can be observed. It is based on a combination of the above-mentioned risk factors. As most of the risk factors associate with ACPA-positive RA, it is suggested that ACPA-positive and ACPA-negative RA are two different subtypes of RA. ACPA-positive RA development is determined by a characteristic genetic background (e.g. *HLA-DRB1*, *PTPN22* and *PADI4*) upon which environmental factors operate to result in synovial inflammation. It is proposed that the environmental factors act in the epithelial layer of the skin, gut and lungs. For example, smoking induces the expression of PAD enzymes and citrullination in the lungs[238]. In addition, the bacterium *P. gingivalis*, which is common in periodontal disease, expresses PAD enzymes and can induce citrullination[239]. These factors can promote ACPA generation. Circulating ACPAs have been found to be present prior to RA onset (up to at least 10 years)[240]. However, the presence of ACPAs

is not sufficient to cause RA. The additional trigger that is required is still to be discovered.

T cells play an important role in the pathogenesis of RA. The association of *HLA-DRB1* SE alleles with RA has pointed to the involvement of CD4+ T cells. Many other RA-associated loci have important roles in T cells. In addition, it has been demonstrated that CD4+ T cells infiltrate the synovial membrane of inflamed joints. Furthermore, a therapy that prevents the interaction of the costimulatory molecules (abatacept) has beneficial effects in RA patients[241,242]. Moreover, CD4+ T cells were shown to be required for disease initiation in an antigen-induced arthritis model[243]. There is also evidence that differentiated CD4+ T cells play a role in the pathogenesis of RA. Th17 cells are present in peripheral blood and synovial fluid of RA patients. These Th17 cells primarily secrete IL-17, which is suggested to promote bone erosion[244]. Regulatory T cells can also be detected in the inflamed joints of RA patients, however, it is suggested that the suppressive function of these cells is impaired by the inflammatory milieu[245]. The importance of regulatory T cells is also highlighted by the enrichment of RA risk loci with active H3K4me3 marks in regulatory T cells[167,246]. CD8+ T cells might play a role in RA as well, indicated by the association of MHC class I SNPs with susceptibility to RA[171]. Moreover, CD4+CD28null T cells with cytotoxic characteristics have been identified in a subset of RA patients[247,248].

1.4.2 Idiopathic inflammatory myopathies

Idiopathic inflammatory myopathy, also known as myositis, is a rare (prevalence ranging from 8.7 to 25 per 100,000[249-251]) heterogeneous group of chronic inflammatory diseases that mainly affect the skeletal muscle. It leads to muscle weakness and is characterized by the presence of inflammatory cell infiltrates, including CD4+ and CD8+ T cells, in muscle tissue[252,253]. These T cells are predominantly of the CD28null phenotypes and might play a role in the destruction of muscle fibers by their cytotoxic properties[254]. Furthermore, myositis is also characterized by the presence of autoantibodies and the dysregulation of type I interferons. Moreover, other organs, such as the skin, joints, heart and lung can be involved as well[255]. Myositis is classified on the basis of distinct clinical and laboratory characteristics into subtypes, including the main subtypes dermatomyositis (DM), polymyositis (PM) and sporadic inclusion body myositis[256-258]. There is currently no cure for myositis, however, different treatments exist to slowdown disease progression, improve muscle function, and prevent damage of other organs.

Myositis can be accompanied by the presence of autoantibodies. Up to 56% of the myositis patients have circulating autoantibodies[259]. These autoantibodies can be classified into myositis associated autoantibodies (MAAs) and myositis specific autoantibodies (MSAs). MAAs are present in other connective tissue diseases and target autoantigens such as Ro60/SSA, Ro52/TRIM21, PM-Scl, U1-RNP and Ku. MSAs are highly selective autoantibodies that, among many others, target the cytoplasmic enzymes aminoacyl-tRNA synthetases (ARSs), which catalyze the ATP-dependent binding of an amino acid to its specific tRNA during protein synthesis[260]. Until now, antibodies against eight different ARSs have been identified and the most common one is anti-Jo-1. Up to 30% of the DM and PM patients have anti-Jo-1

antibodies[259,261]. These autoantibodies may be used as disease markers as they seem to correlate with distinct clinical features. For example, anti-Jo-1 antibodies are associated with ILD in myositis and correlate with disease activity[262].

Dysregulation of type I interferons is another characteristic of myositis. Expression levels of interferon regulated genes have been found to be substantially higher in muscle tissue of DM patients compared to patients with other types of myositis[263]. In addition to muscle tissue, interferon regulated genes have been found to be higher expressed in peripheral blood samples of patients with DM and PM and are correlated with disease activity[264].

The etiology of myositis is unknown, but many studies support the importance of genetic and environmental factors. *HLA-DRB1**03:01 and *HLA-B**08:01 alleles have been determined to be the strongest genetic risk factors for myositis[184,265-267]. However, different alleles might be associated with DM and PM[184,267]. In addition to the *HLA* alleles, *PTPN22* is another strong genetic risk factor for PM[184]. Other suggestive myositis risk loci include *BLK*, *CD28* and *STAT4*[184]. Environmental factors, such as exposure to ultraviolet light[268], vitamin D deficiency[269] and smoking[270] have been found to be associated with myositis. An interaction between smoking and *HLA-DRB1**03 is observed in anti-Jo-1 positive myositis patients[270]. It is suggested that this interaction may prime the development of anti-Jo-1 antibodies.

1.4.2.1 Dermatomyositis

DM typically presents as muscle weakness that is accompanied by skin rash. There is a higher prevalence among older individuals with a peak at 50 - 59 years of age and women are more often affected than men. DM may also be part of other syndromes, including anti-synthetase syndrome and myositis with overlap syndrome[255]. In addition, DM may be associated with cancer. It is reported that up to one third of the patients develop cancer within 5 years after DM diagnosis[271]. Regarding the inflammatory infiltrates in muscle biopsies of DM patients, they are located mainly in the perimysium and in perivascular areas and are predominated by CD4+ T cells with occasional pDCs and B cells[252,253].

1.4.2.2 Polymyositis

PM patients show proximal symmetric weakness as DM patients but without skin rash. It is also more common in women than in men and is usually not seen in the childhood[255]. The cellular infiltrates are located mainly in the endomysium surrounding muscle fibers and typically dominated by CD8+ T cells[252,253]. Perforin in these CD8+ T cells is located towards the target muscle fiber[272] and is suggested to contribute to the muscle cell damage only in PM patients[273], indicating that different mechanisms in DM and PM are involved.

1.4.3 Unmet needs in autoimmune diseases

Autoimmune diseases are a significant clinical problem because of their chronic nature and the associated healthcare cost. Over the past decades, there have been major breakthroughs in understanding the mechanisms of RA and myositis, which resulted in

significant therapeutic advances to manage the symptoms of these disease. However, a clear understanding of the mechanisms that underlie RA and myositis is still lacking, which obstructs the development of remedies for these diseases. In addition, only a few molecular biomarkers have been identified to aid in stratifying RA and myositis patients for identifying the right therapeutics.

2 AIMS

The autoimmune diseases RA and myositis are believed to arise from a combination of genetic and environmental factors. However, a clear understanding of the mechanisms that underlie these diseases is still lacking. The overall aim of the work presented in this thesis is therefore to better understand the functional role of genetic risk factors in the development of RA and myositis.

The specific aims listed according to each paper are:

I. To investigate the functional consequences of autoimmunity associated SNPs in the *PTPN2* locus at chromosome 18.

II. To determine the impact of the *PTPN22* risk allele on CD4+ T-cell function in healthy individuals and RA patients.

III. To find differentially expressed genes in healthy individuals with and without *HLA-DRB1**04:01 alleles, the major genetic risk factor for RA.

IV. To compare whole-genome transcriptomes of CD4+ and CD8+ T cells from peripheral blood between PM and DM.

3 EXPERIMENTAL PROCEDURES

3.1 COHORTS AND METHODOLOGY

In this work, different materials were used. In Paper I and II, data from the Swedish Epidemiological Investigation of Rheumatoid Arthritis (EIRA) study and the COMBINE study was used. In addition, different samples of healthy individuals and patients were collected. For each cohort, the cell sorting and RNA-seq methodology is briefly described. Detailed descriptions of the cohorts and methods can be found in the respective papers.

3.1.1 EIRA study (Paper I)

EIRA is a population-based case-control study. Incident cases of RA, from clinics in the middle and southern region of Sweden, were invited for the study from the year 1996. All RA patients correspond to the 1987 ACR criteria[274]. To each RA patient, at least one healthy individual was randomly selected matched by age, sex and residence area. All participants were asked to fill out a comprehensive questionnaire about their lifestyle. All participants gave informed consent and the study was approved by the ethical review board at the Karolinska Institutet and the Stockholm regional ethical review board.

DNA was extracted from whole blood samples by using a salting-out method and genotyped using the Illumina HumanHap300 Array. SNP imputation was done based on the 1,000 genomes project. DNA methylation levels were profiled using the Illumina HumanMethylation450 BeadChip[275]. In total, data of 335 healthy individuals and 354 ACPA-positive RA patients was available.

3.1.1.1 mQTLs

Genotypes of 37 SNPs (autoimmune associated SNPs in high LD with these variants ($r^2 > 0.8$)) and DNA methylation levels at 417 CpG sites (all CpG sites located in a 2 Mb window around *PTPN2*) were extracted. Associations of these SNPs and CpG methylation levels were calculated using a linear regression model (GEM package). Sex, age and smoking status were included as covariates.

3.1.2 COMBINE study (Paper I and II)

The COMBINE study consists of a large collection of multi-omics data from RA patients. RA patients visiting the Karolinska University Hospital were included in the study and divided into three groups (all patients corresponded to the 1987 ACR or the 2010 ACR/EULAR criteria). Group A includes early RA patients initiating methotrexate (MTX) treatment (naïve to treatment at baseline visit and after continuous treatment with MTX at follow-up visit (approximately three months later)). Group B includes RA patients initiating anti-TNF therapy (failed MTX treatment at baseline visit and after continuous treatment with anti-TNF at the follow-up visit). Group C includes RA patients initiating a second-line biologic treatment (failed anti-TNF therapy at baseline visit and after continuous treatment with second biologic at the follow-up

visit)[276]. In addition, healthy individuals, matched by age and sex, were recruited from the Uppsala Bioresource. All participants gave informed consent and the study was approved by the Stockholm and Uppsala regional ethics committees.

DNA was extracted and genotyped using the Illumina OmniExpress array. SNP imputation was done based on the 1,000 genomes project. RNA was purified from peripheral blood mononuclear cell (PBMC) samples, prepared into sequencing libraries using the Illumina TruSeq RNA library prep kit and sequenced on the Illumina HiSeq 2000. In Paper I, the raw reads were filtered based on quality and adapters were trimmed. The reads were aligned to the hg19 assembly using TopHat2[277]. Gene expression was quantified using HTSeq followed by TMM-normalization, mean-scaling and log2 transformation (edgeR package)[278]. In Paper II, the reads were aligned to the hg38 assembly using STAR[279]. Expression values were determined using rpkmforgenes.py[280]. In total, data of 59 healthy individuals and 137 RA patients was available.

3.1.2.1 eQTLs

Genotypes of 41 SNPs (autoimmune associated SNPs in high LD with these variants ($r^2 > 0.8$)) and expression levels of genes located in a 2 Mb window around *PTPN2* were extracted. Associations between SNPs and gene expression levels were calculated using a linear mixed model (nlme package). For healthy individuals, sex was used as a fixed effect and study individuals' ID as a random effect. For RA patients, sex and patient groups (group A, B and C as described above) were used as fixed effects and study individuals' ID as a random effect.

3.1.3 Healthy individuals – *HLA-DRB1* (Paper I and III)

Blood samples from 32 healthy individuals (females between 55 and 73 years of age), selected by positivity for the *HLA-DRB1* alleles *04, *13:01 and *15:01, were provided by the Uppsala Bioresource. All participants gave informed consent and the study was approved by the Uppsala regional ethics committee.

The blood samples were processed by density gradient centrifugation using Ficoll to recover PBMCs. From the PBMCs, CD4+ T cells, CD8+ T cells, CD14+ monocytes and CD19+ B cells were isolated via positive selection using CD4, CD8, CD14 or CD19 Microbeads (Miltenyi) on the Miltenyi autoMACS® Pro Separator. Total RNA was extracted from the PBMCs and the isolated cell subsets and RNA quantity/quality was assessed. Sequencing libraries were prepared for 90 samples using the Illumina TruSeq stranded total RNA library prep kit with Ribo-Zero and analyzed on an Illumina HiSeq 2500 sequencer. The raw reads were filtered based on quality and adapters were trimmed. Filtered reads were aligned to the hg38 assembly using STAR[279] in two-pass mode with default settings.

3.1.3.1 Mapping to the MHC region

The hg38 assembly includes only one haplotype sequence for the MHC region (PGF, *HLA-DRB1**15:01). Using the above-mentioned standard mapping method, many reads will misalign to the MHC region (chromosome 6: 28,500,000 – 33,500,000).

Therefore, reads were aligned to the eight available MHC reference haplotypes[281]. In short, unmapped reads and reads mapped to the MHC region (using the standard method) were extracted and realigned to the eight MHC reference haplotypes (Table 1) independently using STAR[279] in two-pass mode. AltHapAlignR[282] was used to estimate expression of genes in the MHC region using alignments to the eight available MHC reference haplotypes. Differentially expressed genes were identified using DESeq2[283]. The covariates age (< 65 and ≥ 65), RIN value (groups) and RNA concentration (groups) were used. Differences in gene expression with Benjamini-Hochberg adjusted *P*-value < 0.05 and fold change (log2) > 1 were considered significant.

Table 1. The eight available MHC reference haplotypes showing the MHC class II gene types.

Haplotype	*HLA-DQA1*	*HLA-DQB1*	*HLA-DRB1*
PGF	*010201	*0602	*150101
COX	*050101	*020101	*030101
QBL	*050101	*020101	*030101
APD	-	-	-
DBB	*0201	*030302	*070101
MANN	*0201	*0202	*070101
MCF	*0303	*030101	-
SSTO	*030101	*030501	*040301

Dashes indicate the absence of a gene in the available reference haplotypes.

3.1.4 Healthy individuals – *PTPN22* (Paper I and II)

Blood samples from 26 healthy individuals paired by age and sex, selected on *PTPN22* homozygous genotypes (risk genotype 1858TT and non-risk genotype 1858CC), were provided by the Uppsala Bioresource. All participants gave informed consent and the study was approved by the regional ethical review board in Uppsala.

The blood samples were processed by density gradient centrifugation using Ficoll to recover PBMCs. PBMCs were labeled with LIVE/DEAD Near-IR Dead Cell dye (Invitrogen) and fluorescently labeled with anti-CD14, anti-CD4, anti-CD8, anti-CCR7 and anti-CD45RA to sort naïve CD4+ T cells (CD45RA+CCR7+) using an influx sorter (BD). The naïve CD4+ T cells were activated with anti-CD3/CD28 beads (Dynabeads) for 16 h. Total RNA was extracted of naïve CD4+ T cells before and after activation and RNA quantity/quality was assessed. Sequencing libraries were prepared for 32 samples using the Illumina TruSeq stranded mRNA kit and analyzed on an Illumina HiSeq 2500 sequencer. The reads were aligned to the hg38 assembly using STAR[279] with default settings. Expression values were determined using rpkmforgenes.py[280]. Genes were tested for differential expression using the Mann-Whitney U test. Differences in gene expression with *P*-value < 0.05 and fold change (log2) > 1.2 were considered significant.

3.1.5 Myositis patients (Paper IV)

Blood samples from 33 myositis patients (treated with different therapies) visiting the Karolinska University Hospital were selected for the study on basis of diagnosis (PM and DM) and *HLA-DRB1**03 status (positive and negative). All patients gave informed consent and the study was approved by the Stockholm regional ethics board.

The blood samples were processed by density gradient centrifugation using Ficoll to recover PBMCs. From the PBMCs, CD4+ T cells, CD8+ T cells, CD14+ monocytes and CD19+ B cells were isolated via positive selection using CD4, CD8, CD14 or CD19 Microbeads (Miltenyi) on the Miltenyi autoMACS® Pro Separator. Total RNA was extracted from the isolated cell subsets and RNA quantity/quality was assessed. Sequencing libraries were prepared for 24 samples using the Illumina TruSeq stranded total RNA library prep kit with Ribo-Zero and analyzed on an Illumina HiSeq 2500 sequencer. The raw reads were filtered based on quality and adapters were trimmed. Filtered reads were aligned to the hg38 assembly using STAR[279] with default settings. Differentially expressed genes were identified using DESeq2[283]. The covariates sex, age (< 60 and ≥ 60), RIN value and *HLA-DRB1**03 status (or diagnosis for *HLA-DRB1**03-positive versus *HLA-DRB1**03-negative) were used. Differences in gene expression with Benjamini-Hochberg adjusted P-value < 0.05 were considered significant.

3.2 ADDITIONAL METHODS

In addition to the above-mentioned methods for each specific cohort, more general methods were used, which are briefly described in this section. Detailed descriptions of these methods can be found in the respective papers.

3.2.1 Knockdown of *LINC01882* (Paper I)

Jurkat T cells (LGC Standards) were kept in RPMI-1640 medium supplemented with 10% fetal bovine serum, L-glutamine and penicillin/streptomycin in a 37°C incubator with 5% CO_2. Jurkat T cells were transfected with a mixture of a dsiRNA (50nM; Integrated DNA Technologies) and an antisense LNA™ GapmeR (50nM; Exiqon) targeting the second exon of *LINC01882* using Lipofectamine 2000 according to the manufacturer's protocol. In addition, a mixture of the dsiRNA NC1 (50nM; Integrated DNA Technologies) and an antisense LNA™ GapmeR (50nM; Exiqon) was used as negative control. Cells were harvested 24 and 48 hours after transfection. Total RNA was extracted and RNA quantity/quality was assessed. Sequencing libraries were prepared for eight samples using the Illumina TruSeq stranded total RNA library prep kit with Ribo-Zero and analyzed on an Illumina HiSeq 2500 sequencer. The raw reads were filtered based on quality and adapters were trimmed. Filtered reads were aligned to the hg38 assembly using STAR[279] with default settings (for lncRNAs, the FANTOM CAT assembly was used[64]). Differentially expressed genes, adjusted for time point and experiment, were identified using DESeq2[283]. Differences in gene expression with Benjamini-Hochberg adjusted P-value < 0.05 were considered significant.

3.2.2 Quantitative real-time PCR (Paper I and II)

Efficiency of *LINC01882* knockdown in Jurkat T cells was evaluated by quantitative real-time PCR using iQ™ SYBR® Green Supermix (Bio-Rad laboratories) on a Bio-Rad CFX384 Real-Time PCR system. *SFRS9* and *HPRT* were used to normalize gene expression levels. In addition, *CFLAR* and *TNFRSF9* expression levels were measured by quantitative real-time PCR using SsoAdvanced™ Universal SYBR® Green

Supermix (Bio-Rad laboratories) on a Bio-Rad CFX384 Real-Time PCR system. *GAPDH* and *B2M* were used to normalize gene expression levels. Data of all experiments was analyzed using the comparative CT method[284].

3.2.3 HLA typing (Paper III and IV)

HLA typing was performed using sequence-specific primers from the HLA-DR low-resolution kit (Olerup SSP) and analyzed by agarose gel electrophoresis[285]. In addition, HLA-DR4 subtyping was performed for *HLA-DRB1**04 positive individuals. An interpretation table was used according to the manufacturer's instructions to determine the specific genotype of each sample. Moreover, seq2HLA[286] was used to impute classical HLA alleles from RNA-seq data.

4 RESULTS AND DISCUSSION

4.1 PAPER I: T CELLS ARE INFLUENCED BY A LONG NON-CODING RNA IN THE AUTOIMMUNE ASSOCIATED *PTPN2* LOCUS

4.1.1 Rationale

Previous GWAS have shown that several non-coding SNPs in the *PTPN2* region are associated with different autoimmune diseases, including RA. The Wellcome Trust Case Control Consortium reported an association of the intergenic SNP rs2542151 with Crohn's disease, T1D and RA[287]. Later studies identified a stronger association for the intronic SNP rs1893217 with Crohn's disease and T1D[198,288]. In addition, the intronic SNPs rs2847297, rs62097857 and rs8083786 were reported to be associated with RA[167,168,289]. PTPN2 has been identified to be an important factor in autoimmunity, as it functions as a negative regulator of TCR and cytokine signaling[209]. However, the functional mechanisms by which autoimmune disease-associated SNPs in the *PTPN2* locus influence PTPN2 are not clear. We therefore set out to investigate the link between non-coding SNPs in the *PTPN2* region and autoimmunity.

4.1.2 QTL analyses in healthy individuals and RA patients

To determine the link between risk variants in the *PTPN2* locus and autoimmunity, mQTL and eQTL analyses were performed. mQTL analyses were performed based on DNA methylation data of the EIRA study. In healthy individuals as well as RA patients, a strong association was detected for rs1893217 (and SNPs in high LD with rs1893217) and DNA methylation levels at four CpG sites in a CpG island 7.5 kb downstream of *PTPN2*. Interestingly, three lncRNAs (*RP11-973H7.1*, *RP11-973H7.5* and *LINC01882*) are located downstream of this CpG island (within 30 kb). We therefore tested for associations between risk variants in the *PTPN2* locus and expression levels of *PTPN2* and the lncRNAs *RP11-973H7.1* and *LINC01882* (*RP11-973H7.5* could not be detected). Using GTEx data and COMBINE data, no significant associations were found for risk variants in the *PTPN2* locus and *PTPN2* expression levels in whole blood and PBMC samples of healthy individuals and RA patients. However, a recent study showed that the rs1893217 risk allele is associated with decreased levels of *PTPN2* in CD4+CD45RO+ T cells of healthy individuals[207]. This discrepancy might be due to the materials (whole blood and PBMCs) we used, as the majority of eQTLs are usually cell type specific. On the other hand, we identified significant associations between risk variants in the *PTPN2* locus and expression levels of the lncRNAs *RP11-973H7.1* and *LINC01882* in whole blood and PBMC samples of healthy individuals and RA patients.

SNPs in high LD with rs12971201 are most significantly associated with the expression of *RP11-973H7.1* in healthy individuals and RA patients. Different SNPs are most significantly associated with the expression of *LINC01882* in healthy individuals and RA patients. The SNP rs11875687 is most significantly associated with the expression of *LINC01882* in healthy individuals, whereas the SNPs in high LD with rs12971201

are most significantly associated with the expression of *LINC01882* in RA patients. Although different SNPs are associated with the expression of *LINC01882* in healthy individuals and RA patients, *LINC01882* is not differentially expressed between healthy individuals and RA patients. In addition, the expression levels of *RP11-973H7.1* were also not different between healthy individuals and RA patients. This might reflect that the expression of these lncRNAs are only important in a specific cell type and that the differences in expression might therefore not be identified in PBMC samples. In addition, we measured expression levels in patients with chronic RA, where the differences in lncRNA expression might have already been normalized by other pathways.

The mQTLs and eQTLs are inversely correlated, which is in line with the traditional view that decreased DNA methylation levels are associated with increased gene expression levels. However, it is not clear from our data whether the change in expression of *RP11-973H7.1* and *LINC01882* is a direct consequence of DNA methylation changes or that other factors are involved. The data sources HaploReg[290] and RegulomeDB[291] provided evidence that the SNPs rs2852151 and rs3826557, in high LD with rs12971201, might affect the binding sites for TFs and DNA-damage repair proteins, such as RAD21, CTCF, STAT3, FOS and STAT1. Further studies are needed to identify if these proteins are involved in the regulation of *RP11-973H7.1* and *LINC01882* expression. Moreover, there is data showing that the promoter region of *PTPN2*, where the SNP rs4797709 (in high LD with rs12971201) is located, physically interacts with *LINC01882* in different cell types[292].

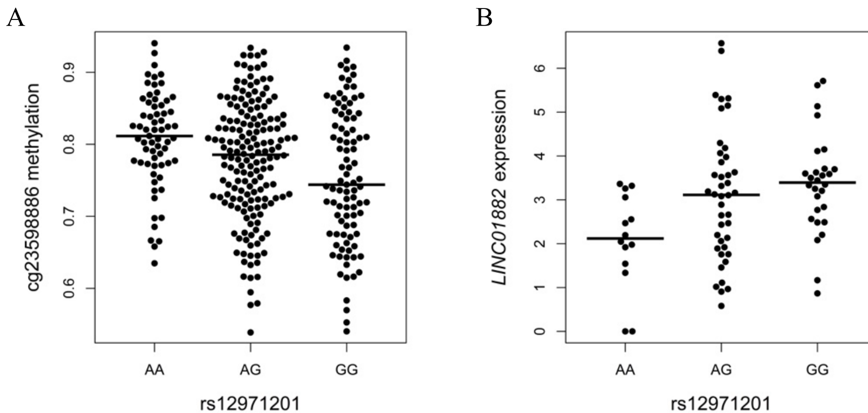

Figure 3. Genotype effect on methylation and expression level in healthy individuals. A) Association between rs12971201 genotype and methylation levels at the CpG site cg23598886 in whole blood of healthy individuals. B) Association between rs12971201 genotype and expression levels of LINC01882 in PBMCs of healthy individuals.

4.1.3 Expression profile of the lncRNAs in T cells

The mQTL and eQTL analyses pointed to the involvement of the lncRNAs *RP11-973H7.1* and *LINC01882* in different autoimmune diseases, however, the function of

these lncRNAs is not known. We determined the cell types in which *RP11-973H7.1* and *LINC01882* are expressed. The lncRNAs were detected in CD4+ T cells and CD8+ T cells, but not in CD14+ monocytes and CD19+ B cells. In addition, a H3K27ac peak just upstream of the lncRNA *RP11-973H7.1* was only identified in different T cell subsets in ChIP-seq data from the Roadmap Epigenomics project. This shows that *RP11-973H7.1* and *LINC01882* are only expressed in different subsets of T cells. Furthermore, we identified that expression of *LINC01882* upon CD3/CD28 stimulation was decreased in naïve CD4+ T cells. Although we identified that the lncRNAs are expressed in T cells, the expression levels are low.

4.1.4 Knockdown of *LINC01882* in Jurkat T cells

To identify potential functions of *LINC01882*, we silenced the expression of this lncRNA in Jurkat T cells using a combination of an antisense oligonucleotide and RNA interference. Silencing of *LINC01882* did not influence the expression levels of *PTPN2*. To identify potential other targets, we performed RNA-seq on these samples. Silencing of *LINC01882* modestly changed the expression of 12 genes, six genes were upregulated and six genes were downregulated. The most upregulated gene upon *LINC01882* knockdown is *BZRAP1*, however, the function of BZRAP1 in immune cells is not known. One of the downregulated genes upon *LINC01882* is *ZEB1*, encoding a TF that represses the levels of *IL-2* when it cooperates with CtBP2[293]. However, it is not predicted that *LINC01882* binds directly to *ZEB1*. It was recently shown that miR-200a-3p negatively regulates the recruitment of ZEB1 and CtBP2 to the promoter of *IL-2*, which promotes the production of IL-2 in T cells[294]. It is predicted that miR-200a-3p and *LINC01882* can interact. We therefore hypothesize that *LINC01882* regulates the expression of *ZEB1* by regulating the levels of miR-200a-3p. Moreover, it is predicted that miR-200a-3p can interact with *KLF12* and *MAP2K4*, which are both affected by the silencing of *LINC01882*. These results suggest that *LINC01882* is potentially involved in the regulation of IL-2 (Figure 4). As mentioned above, IL-2 plays a crucial role in the maintenance of regulatory T cells and dysregulation may lead to the development of autoimmune diseases, including RA.

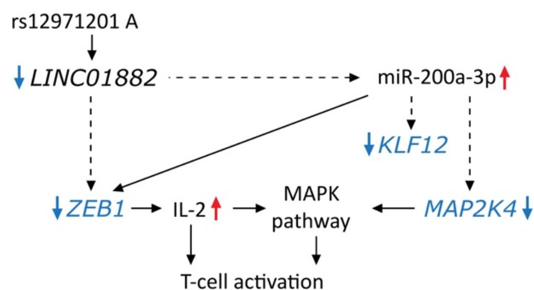

Figure 4. Hypothetical role of LINC01882. SNPs in the PTPN2 locus are associated with the expression of LINC01882. Silencing of LINC01882 decreases the expression of ZEB1, KLF12 and MAP2K4, which might be mediated via the miRNA miR-200a-3p. The protein of ZEB1 represses the expression of IL-2, which is involved in T-cell activation. Blue indicates changes in expression of genes supported by our data, while red indicates suggestive changes.

4.1.5 Concluding remarks

We show that autoimmune disease-associated SNPs in the *PTPN2* locus are associated with the expression of several lncRNAs, including *LINC01882*. We identified that *LINC01882* might play a role in T-cell activation by regulating IL-2 levels. Overall, this suggests that *LINC01882* can be linked to several autoimmune diseases, including RA.

4.2 PAPER II: EOMES-POSITIVE CD4+ T CELLS ARE INCREASED IN PTPN22 (1858T) RISK ALLELE CARRIERS

4.2.1 Rationale

One of the strongest susceptibility genes for many autoimmune diseases, including RA, is *PTPN22*[181-185]. The *PTPN22* risk allele (rs2476601; 1858T) substitutes an arginine (R) to tryptophan (W) in codon 620 (R620W) of the non-catalytic part of LYP (encoded by *PTPN22*). LYP has been identified as a negative regulator of TCR signaling. However, the functional impact of LYP620W is still controversial. Studies in T cells have described that LYP620W could be a gain-of-function variant[193], leading to decreased TCR signaling, or a loss-of-function variant[194], leading to increased TCR signaling. In addition, knock-in mice expressing the LYP variant homolog PEP619W had increased TCR signaling and expansion of effector-memory T cells[195,196]. We therefore set out to study the functional consequences of the *PTPN22* risk allele on CD4+ T-cell function in healthy individuals and RA patients.

4.2.2 Differential expression in naïve CD4+ T cells

As LYP is primarily expressed in T cells[189], we first evaluated if the *PTPN22* risk genotype affects the frequency of CD4+ Th cell subsets in healthy individuals. The frequencies of Th1 (CXCR3+), Th17 (CCR6+), Th1Th17 (CXCR3+CCR6+) and Tfh (CXCR5+) cell subsets were not different between healthy individuals carrying 1858CC and 1858TT. In addition, there were no differences observed in the frequencies of total regulatory CD4+ T cells, resting regulatory T cells (CD45RA+FOXP3^{low}) and activated regulatory T cells (CD45RA-FOXP3^{high}) between healthy individuals carrying 1858CC and 1858TT. A trend was observed towards an increase of the non-regulatory T cell fraction (CD45RA-FOXP3^{low}) in 1858TT carriers.

Although no significant differences were observed in the frequency of the T-cell subsets between 1858CC and 1858TT carriers, it is possible that the small differences that were observed could affect our further analyses. We therefore isolated a homogenous population of naïve CD4+ T cells (CD45RA+CCR7+) from healthy individuals carrying 1858CC and 1858TT. TCR activation of the isolated naïve CD4+ T cells modified the expression of a large number of genes, however, most of these genes were not differentially expressed between 1858CC and 1858TT carriers. This shows that the presence of the risk allele does not induce a global change of TCR-regulated genes in naïve CD4+ T cells. This was further confirmed by no observed differences in calcium flux in CD4+ T cells from healthy individuals carrying the 1858CC and 1858TT genotype.

PTPN22 was not differentially expressed between 1858CC and 1858TT carriers before and after TCR activation. However, 57 genes were differentially expressed between 1858CC and 1858TT carriers after TCR activation. Among these genes, 20 genes were higher expressed in healthy individuals carrying 1858CC and 37 genes were higher expressed in healthy individuals carrying 1858TT. The genes higher expressed in 1858TT carriers were enriched with genes belonging to the Gene Ontology terms[295] ribosome and apoptosis. It has been shown that TCR stimulation increases the expression of ribosomal proteins[296]. Higher levels of ribosomal protein genes might suggest that *PTPN22* risk allele carriers have increased cell proliferation. This is further supported by higher expression levels of *CFLAR* and *TNFRSF9* in 1858TT carriers (Figure 5). These results were confirmed by quantitative real-time PCR. *CFLAR* encodes the protein FLIP, which suppresses apoptosis by forming heterodimers with pro-caspase 8 to inhibit its activation[297]. It has been shown that FLIP-deficient T cells have defective survival upon TCR stimulation[298]. In addition, high levels of FLIP have been found in cytotoxic CD4+CD28null T cells[299,300]. *TNFRSF9* encodes the protein 4-1BB (CD137), which is a co-stimulatory molecule transiently expressed after TCR stimulation[301]. We identified that the expression of 4-1BB on the surface of CD4+ T cells stimulated with staphylococcal enterotoxin B for four days was increased in ten out of twelve individuals carrying the *PTPN22* risk genotype. Co-stimulation through 4-1BB induces the expression of the TF EOMES[302,303], which drives cytotoxic differentiation. Since the expression of *EOMES* was not induced in naïve CD4+ T cells after TCR activation, we did not observe differential expression of *EOMES* in naïve CD4+ T cells of healthy individuals carrying 1858CC versus 1858TT. However, higher expression levels of *EOMES* were observed in PBMCs of healthy individuals carrying 1858CT versus 1858CC from the COMBINE study (Figure 6A). Moreover, *CFLAR* and *TNFRSF9* have been found to be associated with RA[167]. Overall, this suggests that CD4+ T cells of healthy individuals carrying the *PTPN22* risk allele are more resistant to apoptosis and more prone to differentiate into cytotoxic T cells.

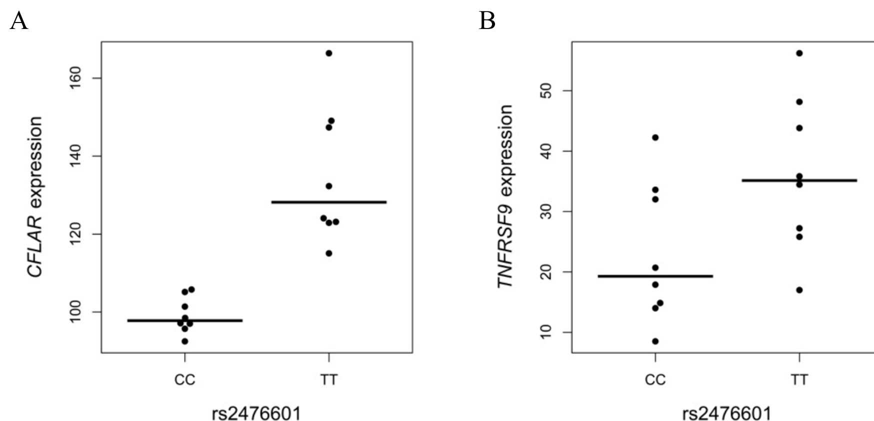

Figure 5. Differential expression of CFLAR and TNFRSF9 in TCR stimulated naïve CD4+ T cells of healthy individuals carrying the PTPN22 risk genotype (TT) and non-risk genotype (CC). A) Differential expression of CFLAR in 1858CC and 1858TT carriers. B) Differential expression of TNFRSF9 in 1858CC and 1858TT carriers.

4.2.3 EOMES+CD4+ T cells in healthy individuals and RA patients

These results led us to examine the frequency of EOMES+CD4+ T cells in PBMCs of healthy individuals carrying 1858CC and 1858TT. We found that the frequency of EOMES+CD4+ T cells was increased in healthy individuals carrying the *PTPN22* risk genotype (Figure 6B). As the expression of EOMES in naïve CD4+ T cells is virtually absent[304], the increased expression of EOMES in CD4+ T cells could be due to a skewed frequency of naïve CD4+ T cells in healthy individuals carrying 1858TT. We indeed found that nine out of thirteen healthy individuals carrying 1858TT presented a decreased frequency of naïve CD4+ T cells. The frequency of EOMES tended to be increased in memory CD4+ T cells (CCR7-CD45RA-) of healthy individuals carrying 1858TT. However, the frequency of EOMES was not affected in the different memory CD4+ T-cell subsets, including central memory (CCR7+CD45RA-), effector memory (CCR7-CD45RA-) and terminally differentiated effector memory (CCR7-CD45RA+) cells. Thus we showed that healthy individuals carrying the *PTPN22* risk genotype have a higher frequency of EOMES+CD4+ T cells, which is correlated with a decreased frequency of naïve CD4+ T cells.

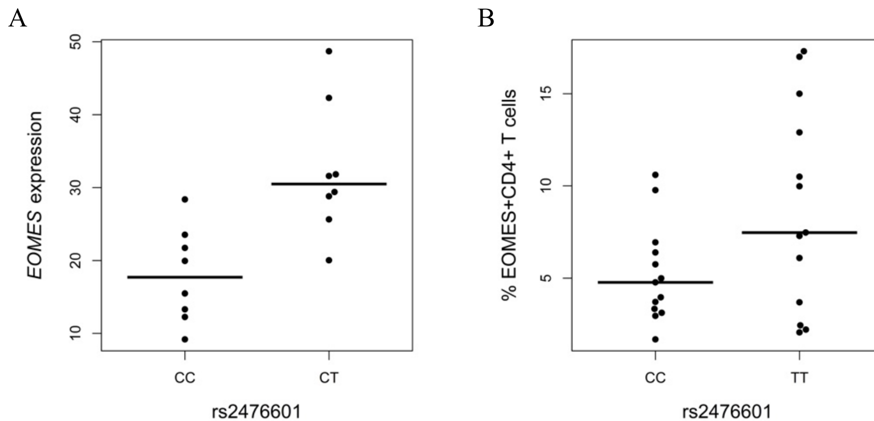

Figure 6. RNA and protein levels of EOMES in healthy individuals carrying the PTPN22 risk allele (1858T) A) Differential expression of EOMES in PBMCs of 1858CC and 1858CT carriers. B) Frequency of EOMES+CD4+ T cells in 1858CC and 1858TT carriers.

We then investigated the relevance of EOMES+CD4+ T cells in RA patients. We collected paired samples of peripheral blood and synovial fluid from twelve RA patients with different *PTPN22* genotypes. We found that the frequency of EOMES+CD4+ T cells was significantly higher in synovial fluid compared to peripheral blood, as all CD4+ T cells are of memory phenotype in synovial fluid. However, the frequency of EOMES+CD4+ T cells was not different in synovial fluid of RA patients carrying the different *PTPN22* genotypes. This might suggest that other factors in the synovial fluid of RA patients can contribute to their differentiation. Moreover, differences in treatment and inflammatory processes between the RA patients might contribute to mask part of the effect of the risk allele in RA patients. Furthermore, EOMES has been shown to regulate the levels of perforin 1[305,306]. We

found an increased frequency of perforin 1 in EOMES+CD4+ T cells compared to EOMES-CD4+ T cells, with higher frequencies observed in RA patients carrying the *PTPN22* risk allele. In addition, perforin 1 expression is correlated with granzyme A expression and to a lesser extent with granzyme B expression. Notably, granzyme B expression was restricted to EOMES+CD4+ T cells while granzyme A expression could be found to some extent in EOMES-CD4+ T cells. Moreover, levels of perforin 1 are higher in ACPA-positive RA patients carrying the *PTPN22* risk allele. The same results were observed in EOMES+CD8+ T cells of synovial fluid from RA patients, however, perforin 1 expression was not different between *PTPN22* risk and non-risk alleles.

Although this data demonstrates a link between *PTPN22* risk alleles and the generation of CD4+ T cells with cytotoxic characteristics, the exact mechanisms driving the generation of EOMES+CD4+ T cells in RA and other autoimmune diseases is currently unknown. It has been shown that the expression of soluble 4-1BB and 4-1BBL is increased in RA patients[307], which might contribute to the induction of EOMES. In addition, pro-inflammatory cytokines such as IL-15 could participate in driving the accumulation of cytotoxic CD4+ T cells in synovial tissue of RA patients[308,309].

4.2.4 Concluding remarks

We identified that the frequency of EOMES+CD4+ T cells is increased in healthy individuals carrying the *PTPN22* 1858TT risk genotype compared to healthy individuals carrying the *PTPN22* 1858CC non-risk genotype. Overall, this suggests a new role for the *PTPN22* risk genotype in the context of RA through the generation of CD4+ T cells with cytotoxic characteristics.

4.3 PAPER III: RNA EXPRESSION OF HLA-DRB AND HLA-DQ GENES IN HEALTHY INDIVIDUALS DIFFERS BETWEEN MHC CLASS II HAPLOTYPES

4.3.1 Rationale

The MHC class II gene *HLA-DRB1* is the major genetic susceptibility locus for RA and many other autoimmune diseases. *HLA-DRB1* SE alleles have been associated to ACPA-positive RA[171] and are proposed to reflect a favored binding of citrullinated peptides to the binding groove of the HLA-DR beta chain[172]. However, other mechanisms such as differential gene expression linked to certain *HLA-DRB1* alleles have not been thoroughly investigated. We therefore studied genome-wide gene expression profiles in PBMCs, CD4+ T cells, CD8+ T cells and CD14+ monocytes of healthy individuals with *HLA-DRB1* SE-positive (*04) and SE-negative (*13:01 or *15:01) alleles. *HLA-DRB1**13:01 alleles provide protection against ACPA-positive RA[176], while *HLA-DRB1**15:01 alleles are not associated with RA but confer the strongest risk for developing other autoimmune diseases, including MS[310].

4.3.2 Differential expression in PBMCs

We first focused on identifying differentially expressed genes in PBMCs of healthy individuals carrying *HLA-DRB1* SE-positive versus SE-negative alleles. We identified no differentially expressed non-MHC genes between *HLA-DRB1* SE-positive and SE-negative individuals (adjusted *P*-value < 0.05 and fold change (log2) > 1). It might be possible that cell type specific differences in gene expression were missed due to the heterogenous mixture of cell types in the PBMC samples.

Using the eight MHC reference haplotypes for mapping reads to the MHC region (described above), we identified five MHC class II genes that were differentially expressed between *HLA-DRB1* SE-positive and SE-negative individuals. We found that *HLA-DRB4*, *HLA-DQA2*, *HLA-DRB1* and *HLA-DQA1* had higher expression levels in *HLA-DRB1* SE-positive individuals, whereas *HLA-DQB1* had higher expression levels in *HLA-DRB1* SE-negative individuals. The differential expression of *HLA-DRB4* between *HLA-DRB1* SE-positive and SE-negative individuals is expected as we compare a group that has *HLA-DRB4* (individuals carrying *HLA-DRB1**04) with a group that does not have *HLA-DRB4* (individuals carrying *HLA-DRB1**13 and *15 (Figure 7)). Interestingly, the expression of *HLA-DRB1* was also different between *HLA-DRB1* SE-positive and SE-negative individuals, although *HLA-DRB1* is present in all groups (individuals carrying *HLA-DRB1**04, *13 and *15 (Figure 7)). In addition to *HLA-DRB1*, the expression of *HLA-DQA1*, *HLA-DQA2* and *HLA-DQB1* was different between *HLA-DRB1* SE-positive and SE-negative individuals, which is consistent with the extended HLA-DR/HLA-DQ haplotype.

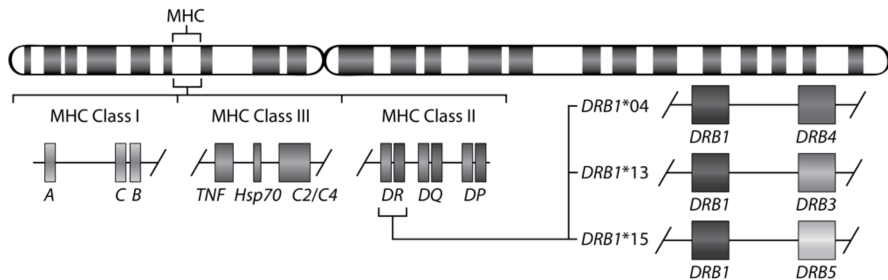

*Figure 7. Gene map of the MHC region. The MHC region on the short arm of chromosome 6 contains MHC class I, II and III. MHC class II genes include HLA-DR, HLA-DQ and HLA-DP. The different HLA-DR haplotypes (HLA-DRB1*04, *13 and *15) are linked with only one of the following genes: HLA-DRB4, HLA-DRB3 or HLA-DRB5.*

Although this strategy mapped the reads more accurately[282], there are several issues that remain and may interfere with gene expression profiling in the MHC locus. First, the amount of available MHC reference haplotypes is limited[281]. There are only references available for the haplotypes with *HLA-DRB1**03:01, *04:03, *07:01 and *15:01 alleles (Table 1). As there is no MHC reference haplotype containing *HLA-DRB1**13:01, we had to remove all individuals carrying *HLA-DRB1**13:01 from our analysis. Second, the *HLA-DRB1**04:01 samples are mapping to the best matched MHC reference haplotype SSTO (*HLA-DRB1**04:03), which could potentially lower the number of

mapped reads. Moreover, the efficiency of read mapping can be different for different haplotypes. We therefore performed our differential expression analyses on samples with only *HLA-DRB1**03:01 as second allele.

Taken these issues into account, we identified three MHC class II genes that were differentially expressed between individuals carrying *HLA-DRB1**03:01/*04:01 and *03:01/*15:01. As expected, *HLA-DRB4* was higher expressed in *HLA-DRB1**03:01/*04:01 individuals and *HLA-DRB5* was higher expressed in *HLA-DRB1**03:01/*15:01 individuals. In addition, *HLA-DQB1* was higher expressed in *HLA-DRB1**03:01/*15:01 individuals compared to *HLA-DRB1**03:01/*04:01 individuals. Although *HLA-DRB1* was higher expressed in individuals carrying *HLA-DRB1**03:01/*04:01 compared to *HLA-DRB1**03:01/*15:01, it did not reach statistical significance due to the small sample size. These results show that *HLA-DRB* and *HLA-DQ* genes are differentially expressed in PBMCs of *HLA-DRB1**03:01/*04:01 versus *03:01/*15:01 healthy individuals.

4.3.3 Differential expression in isolated cell subsets

To identify whether these observed differences were cell type specific, we performed differential expression analyses on isolated CD4+ T cells, CD8+ T cells and CD14+ monocytes. In the CD4+ and CD8+ T-cell subsets, we identified no differentially expressed non-MHC genes between *HLA-DRB1* SE-positive and SE-negative individuals. In CD14+ monocytes, we identified six non-MHC genes that were differentially expressed between *HLA-DRB1* SE-positive and SE-negative individuals. However, there were only ten CD14+ monocyte samples sequenced and the RIN value of most of these samples was relatively low. We therefore suggest that these findings should be first replicated in independent material before any conclusion is drawn.

In CD4+ T cells, we found that the genes *HLA-DRB4*, *HLA-DRB1*, *HLA-DQA2* and *HLA-DQA1* had higher expression levels in healthy individuals carrying *HLA-DRB1* SE-positive alleles compared to SE-negative alleles. In CD8+ T cells, we found that *HLA-DRB4* and *HLA-DRB1* had higher expression levels in *HLA-DRB1* SE-positive individuals, whereas *HLA-DQB1* had higher expression levels in *HLA-DRB1* SE-negative individuals. In CD14+ monocytes, we identified the same differentially expressed genes as in CD4+ T cells. In addition, we found higher expression levels of *HLA-A* in CD14+ monocytes of *HLA-DRB1* SE-positive versus SE-negative individuals. Taken the above-mentioned issues into account, we found the same genes to be differentially expressed in CD4+ and CD8+ T cells of *HLA-DRB1**03:01/*04:01 versus *03:01/*15:01 healthy individuals, although not statistically significant (Figure 8). The small number of CD14+ monocyte samples, however, did not allow us to test differential expression in *HLA-DRB1**03:01/*04:01 versus *03:01/*15:01 individuals.

These results show that *HLA-DRB* and *HLA-DQ* genes are differentially expressed in all tested cell populations of healthy individuals carrying *HLA-DRB1**04:01 versus *15:01 alleles. It has previously been shown that *HLA-DRB1* gene expression levels in peripheral blood B cells were greatly influenced by *HLA-DRB1* alleles[311]. However, the functional consequences of these differences in gene expression are to be discovered. In order to address functional consequences, it is critical to detect protein

levels of these genes as well. In addition, the function of HLA-DR in CD4+ T cells is still debated. Recent studies indicated that pathogenic T cells express HLA-DR in synovial fluid of RA patients[312,313]. In addition, HLA-DR expression on CD4+ T cells has also been proposed to be a marker of regulatory T cells[314]. Although these studies show that there are CD4+ T-cell populations that express HLA-DR, we cannot totally exclude the possibility that our results in CD4+ T cells are a reflection of the contamination by CD14+ monocytes. Moreover, as *HLA-DRB1**04:01 is associated with RA, it will be interesting to measure *HLA-DRB* and *HLA-DQ* levels in sorted cells of RA patients with different *HLA-DRB1* alleles.

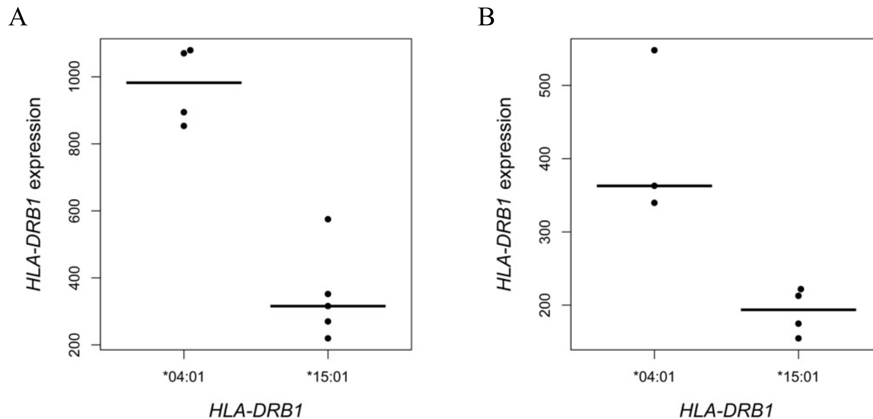

*Figure 8. Differential expression of HLA-DRB1 in isolated cells of HLA-DRB1*04:01 versus *15:01 healthy individuals. A) Differential expression of HLA-DRB1 in CD4+ T cells of HLA-DRB1*04:01 versus *15:01 individuals. B) Differential expression of HLA-DRB1 in CD8+ T cells of HLA-DRB1*04:01 versus *15:01 individuals.*

4.3.4 Concluding remarks

We show that *HLA-DRB1* and *HLA-DQ* genes are differentially expressed in all tested cell populations of healthy individuals carrying *HLA-DRB1**04:01 versus *15:01 alleles. These results suggest that the regulation of *HLA-DRB* and *HLA-DQ* gene expression could represent one of the mechanisms underlying the association of *HLA-DRB1* SE alleles to RA.

4.4 PAPER IV: T-CELL TRANSCRIPTOMICS FROM PERIPHERAL BLOOD HIGHLIGHTS DIFFERENCES BETWEEN POLYMYOSITIS AND DERMATOMYOSITIS

4.4.1 Rationale

PM and DM are chronic inflammatory disorders clinically characterized by skeletal muscle weakness and muscle inflammation[315]. Inflammatory cell infiltrates are commonly found in the muscle of both PM and DM patients[252,316]. In patients with PM, the cellular infiltrates are typically dominated by CD8+ T cells[272,317]. In contrast, in

patients with DM, the cellular infiltrates are predominated by CD4+ T cells with occasional pDCs and B cells[252]. Although the T cell lineage is different between PM and DM, they both display a cytotoxic signature in the absence of the costimulatory CD28 receptor[254,273]. In addition, some of the identified genetic risk factors (e.g. *PTPN22*) are different between PM and DM[184]. These differences suggest that PM and DM are two different disorders. In this study, we addressed whether whole-genome transcriptomes of CD4+ and CD8+ T cells from peripheral blood were different between patients with PM and DM.

4.4.2 Differential expression in CD4+ T cells

We first focused on identifying differentially expressed genes in CD4+ T cells of patients with PM versus patients with DM. We found that the overall expression of genes in CD4+ T cells was very similar between PM and DM patients (examined by principal component analysis (PCA)). Using DESeq2[283], we identified thirtheen genes that were differentially expressed between PM and DM patients. Among these genes, six were higher expressed in PM patients and seven were higher expressed in DM patients. However, the PCA plot indicated three potential outliers. By performing cell type enrichment analysis from gene expression data[318], we found that these three samples have higher numbers of monocytes than the other samples. As described above, this is due to the isolation method (positive selection by CD4) we used. To exclude the possibility that the differentially expressed genes were detected because of a difference in cell composition, we excluded these samples and performed the analysis again. Four genes were found to be differentially expressed between PM and DM patients, two were higher expressed in PM patients and two were higher expressed in DM patients. We attempted to avoid false positive results by considering only genes that were found to be differentially expressed in both analyses (with and without potential outliers). This resulted in two genes, *ANKRD55* and *S100B*, that were found to have higher expression levels in CD4+ T cells of PM patients compared to DM patients. The protein encoded by *ANKRD55* is suggested to be involved in mediating protein-protein interactions[319], however, the exact function is unknown. Interestingly, SNPs in the *ANKRD55* locus have been found to be associated with several autoimmune diseases, including RA, Crohn's disease and MS[167,202,320-322]. The protein encoded by *S100B* is involved in calcium-dependent regulation of a variety of intracellular activities[323]. *S100B* is normally not expressed in CD4+ T cells, but only in CD8+ T cells and NK cells[324]. This, in combination with low expression levels of *S100B*, might suggest either that the expression of *S100B* is evidence of contamination by other cell types or that this expression is a characteristic of CD4+ T cells in myositis patients. Although this data needs to be replicated in an independent group of myositis patients, we find that CD4+ T cell transcriptomes of patients with PM and DM are rather similar.

The major genetic risk factor for myositis is *HLA-DRB1**03. We therefore investigated which genes were differentially expressed in CD4+ T cells of *HLA-DRB1**03-positive and *HLA-DRB1**03-negative myositis patients. These subtypes of myositis patients were also not separated by the first principal components, suggesting that the overall expression of genes in CD4+ T cells was not different between *HLA-DRB1**03-positive and *HLA-DRB1**03-negative myositis patients. After performing both analyses (with

and without potential outliers), we identified six genes that were differentially expressed between *HLA-DRB1**03-positive and *HLA-DRB1**03-negative myositis patients. Among these genes, *PI4KAP1* was found to have higher expression levels in *HLA-DRB1**03-positive myositis patients and *TRGC2*, *CTSW*, *HPCAL4*, *ZNF683* and *GOLGA8B* were found to have higher expression levels in *HLA-DRB1**03-negative myositis patients. The protein encoded by *CTSW* (cathepsin W) is involved in antigen processing and is exclusively expressed in CD8+ T cells and NK cells[325]. The difference in *CTSW* expression levels might be due to contamination with other cell types, however, it might be possible that *CTSW* is expressed in CD4+ T cells of myositis patients but not translated into proteins. The protein encoded by *ZNF683* is involved in the transcriptional regulation of effector functions, such as production of IFN-γ and granzyme B[326,327]. Moreover, ZNF683 has been found to be upregulated in T cells with cytotoxic characteristics[328]. In addition to CD4+ T cells, we found higher levels of *ZNF683* in CD8+ T cells of *HLA-DRB1**03-negative myositis patients compared to *HLA-DRB1**03-positive myositis patients (as discussed in the next section).

4.4.3 Differential expression in CD8+ T cells

We then focused on identifying differentially expressed genes in CD8+ T cells of patients with PM versus patients with DM. For the CD8+ T-cell samples, the patients with PM were all *HLA-DRB1**03-positive and the patients with DM were all *HLA-DRB1**03-negative. The PCA plot showed no clustering of patients with PM and DM, suggesting that the overall gene expression in CD8+ T cells of patients with PM and DM is similar. However, we identified 588 genes that were differentially expressed between patients with PM and DM. Among these genes, 182 were higher expressed in PM patients and 406 were higher expressed in DM patients. We found that one of the CD8+ T cell samples was clustering with the CD4+ T cell samples in the overall analysis. We therefore excluded this sample and performed the differential expression analysis again. We now identified 308 genes that were differentially expressed between patients with PM and DM. Among these genes, 107 were higher expressed in PM patients and 201 were higher expressed in DM patients. By considering only the genes that were differentially expressed in both analyses (with and without potential outlier), we identified 176 genes that were differentially expressed between patients with PM and DM. Among these genes, 44 were higher expressed in PM patients and 132 were higher expressed in DM patients. For these 176 genes, the enriched Gene Ontology biological processes included lymphocyte migration and regulation of T-cell differentiation (Figure 9). In addition, we noted that the expression of two granzyme encoding genes, *GZMH* and *GZMB*, was higher in CD8+ T cells of DM patients compared to PM patients. It has been shown that the secretion of granzyme B by CD28null T cells may cause muscle cell damage[273]. Moreover, granzyme B cleavage sites have been identified in autoantigens, such as FHL1 and HisRS, which are targeted in myositis[329,330].

TGFB1, the transcript of one of the predominantly reported cytokines (TGF-β) in myositis[331], had a higher expression in CD8+ T cells of DM patients compared to PM patients. Moreover, *ZEB2*, which encodes a TF that plays a role in TGF-β signaling, also had a higher expression in DM patients compared to PM patients. It has been found

that ZEB2 promotes CD8+ T-cell differentiation in response to viral infection[332,333]. ZEB2 is able to interact with SMAD proteins, which are the main signal transducers for receptors of TGF-β[334]. In our data, *SMAD7* had a higher expression in DM patients compared to PM patients. Together, this may suggest that TGF-β signaling is upregulated in CD8+ T cells of DM patients compared to PM patients.

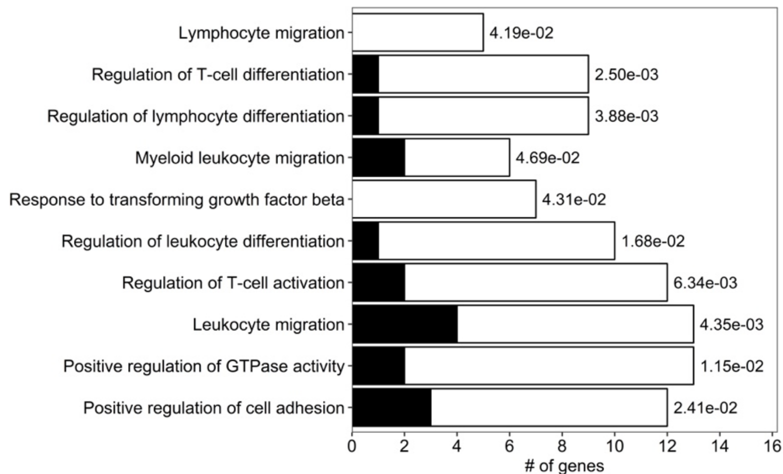

Figure 9. The top 10 biological processes identified by Gene Ontology analysis of the differentially expressed genes in CD8+ T cells of PM and DM patients (sorted by fold enrichment). Bars represent the number of genes higher expressed in patients with PM (black) and higher expressed in patients with DM (white). Fisher's exact test with false discovery rate correction was used to determine significant biological processes.

We also found two TCR beta variable genes, *TRBV28* and *TRBV30*, with higher expression in CD8+ T cells of PM patients compared to DM patients. TRBV28 has been found to be one of the most common TCR variable segments in muscle tissue of myositis patients carrying *HLA-DRB1**03 alleles[335]. As mentioned above, all PM patients are *HLA-DRB1**03-positive and all DM patients are *HLA-DRB1**03-negative, suggesting that these genes are probably differentially expressed due to the HLA status of these patients. In addition, differential expression of these genes might reflect the expansion of pathogenic T-cell clones in this subset of patients.

4.4.4 Concluding remarks

We observed significantly more differentially expressed genes in the CD8+ T-cell subset than in the CD4+ T-cell subset when comparing PM versus DM patients. Although this data needs to be replicated in a bigger independent cohort of patients with PM and DM, it suggests that immune mechanisms related to CD8+ T cells may significantly vary between patients with PM and DM. It also emphasizes that CD8+ T cells are of general interest when studying PM and DM.

5 CONCLUSIONS AND FUTURE PERSPECTIVES

Over the years, many susceptibility loci for autoimmune diseases have been identified, however, the functional mechanisms behind these associations are lacking. As the title 'gene expression profiling in autoimmune diseases' suggests, this thesis aimed on creating a better understanding of the functional role of genetic risk factors in the development of RA and myositis by focusing on alterations in gene expression.

Genetic variants in the *PTPN2* locus have been found to be associated with several autoimmune diseases, including RA. Although PTPN2 might play an important role in autoimmunity, the exact mechanisms by which these genetic variants influence PTPN2 are not clear. In Paper I, we found that genetic variants in the *PTPN2* locus are associated with the expression of the lncRNAs *LINC01882* and *RP11-973H7.1*, but not with the expression of *PTPN2*, in both healthy individuals and RA patients. By silencing *LINC01882* in Jurkat T cells, we suggest that *LINC01882* might play a role in T-cell activation by regulating IL-2 levels, an important cytokine in autoimmunity. This is one of the few examples that shows that genetic variants in intronic regions do not necessarily affect the closest gene. It will be of interest to study the function of the lncRNA *RP11-973H7.*1 in the context of RA as well. Our findings suggest the importance of these lncRNAs in other autoimmune diseases that were previously identified to be associated with genetic variants in the *PTPN2* locus.

In Paper II, we focused on understanding the functional role of a genetic variant (rs2476601) in the coding region of another member of the protein tyrosine phosphatase non-receptor family, namely *PTPN22*. *PTPN22* is recognized as the second most important risk locus (after *HLA-DRB1*) for ACPA-positive RA. The functional impact of this genetic variant is not fully understood, as studies have described contradictory functions. We found that the presence of the *PTPN22* risk genotype does not result in TCR blunting in naïve CD4+ T cells of healthy individuals. Instead, we found specific changes in expression of genes related to T-cell survival (*CFLAR*) and cytotoxic T-cell differentiation (*TNFRSF9*). This led us to identify an increased frequency of EOMES+CD4+ T cells in healthy individuals carrying the *PTPN22* risk genotype. Although the exact mechanism is still to be discovered, we show a new role for the *PTPN22* risk allele in the context of RA through the generation of CD4+ T cells with cytotoxic characteristics. In addition, there are several lncRNAs encoded on the *PTPN22* locus. It will be interesting to study whether the *PTPN22* risk allele is associated with the expression of these lncRNAs. It might be possible that these lncRNAs influence the levels of the TF EOMES in CD4+ T cells of *PTPN22* risk allele carriers.

The MHC class II gene *HLA-DRB1* is the major genetic susceptibility locus for RA. Among others, the *HLA-DRB1**04:01 allele is associated with ACPA-positive RA and is proposed to reflect a favored binding of citrullinated peptides to the binding groove of the HLA-DR beta chain. In Paper III, we studied differences in gene expression between healthy individuals carrying RA related *HLA-DRB1* risk and non-risk alleles.

In PBMCs, CD4+ T cells and CD8+ T cells, we found no differentially expressed non-MHC genes between individuals carrying *HLA-DRB1**04:01 and *15:01 alleles. However, we identified a difference in the expression of *HLA-DRB* and *HLA-DQ* genes between individuals carrying *HLA-DRB1**04:01 and *15:01 alleles. These differences in gene expression were observed in PBMCs, as well as in CD4+ T cells and CD8+ T cells. It will be of interest to study the protein levels of HLA-DR beta chains in individuals carrying different *HLA-DRB1* alleles. This data suggests that *HLA-DRB1* and *HLA-DQ* levels, and potentially their corresponding proteins, might support loss of immune tolerance in RA patients carrying *HLA-DRB1**04:01 alleles.

In Paper IV, we studied differentially expressed genes between patients with PM and DM to differentiate involvement of CD4+ and CD8+ T cells in these myositis subgroups. In CD4+ T cells, we identified that the genes *ANKRD55* and *S100B* had higher expression levels in patients with PM than in patients with DM. In CD8+ T cells, 176 genes were differentially expressed in patients with PM compared to patients with DM. Although these results need to be confirmed in a larger group of patients, several differentially expressed genes between patients with PM and DM were related to cytotoxic T cells. These results add to the evidence that different immune mechanisms are involved in patients with PM compared to patients with DM. Interestingly, the genetic variant in the *PTPN22* locus is also associated with PM but not with DM. Therefore, it will be interesting to study EOMES+CD4+ T cells in patients with PM and DM.

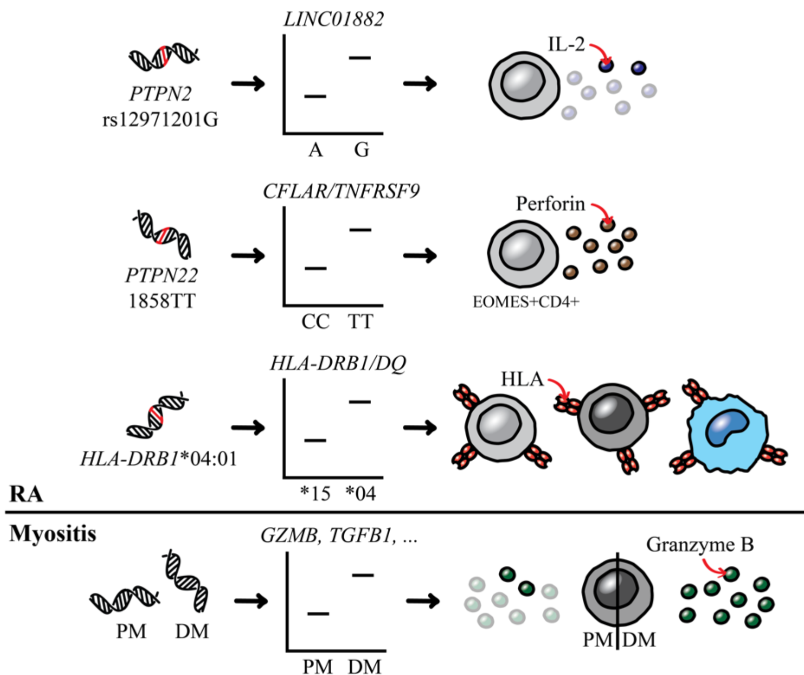

Figure 10. Schematic overview of potential immune mechanisms in subgroups of RA and myositis patients.

My studies highlight the importance of using specific cell subsets instead of whole blood to study the mechanisms of disease-associated genetic variants. However, most of the cell types we studied were obtained from peripheral blood, which might not exactly reflect the cell types in the disease affected tissues. In addition, we collected for the first three studies healthy individuals with certain risk alleles to be able to separate genotype effects from phenotype effects. This resulted in the identification of new molecular pathways in RA patients with certain risk alleles. Finding the exact factors that drive the process of cytotoxic CD4+ T cell proliferation and differential expression of *HLA-DRB1* and *HLA-DQ* genes is expected to lead to the identification of new therapeutic targets for personalized medicine. Promising therapeutic targets might be disease-related lncRNAs, such as *LINC01882*. In addition, the increasing evidence that different immune mechanisms are involved in subgroups of myositis patients, highlight the need for specific treatments for PM and DM. Further studies are needed to conclude if some of the identified differentially expressed genes might be useful for therapeutic development.

Overall, we identified new molecular pathways for genetic risk variants in the *PTPN2*, *PTPN22* and *HLA-DRB1* loci in the context of RA and other autoimmune diseases. We suggest that different immune mechanisms are involved in subgroups of RA and myositis patients (Figure 10). These results could ultimately lead to the identification of new therapeutic targets for different autoimmune diseases.

6 ACKNOWLEDGEMENTS

Over the course of my PhD studies, I have had the chance to meet many people with varying personalities and thus different views on science. All these people have in some way directly or indirectly contributed to the final state of this thesis, for which I am grateful. Therefore, I would like to dedicate this chapter to thank those mentioned below, but also those that have not been explicitly mentioned here… Bedankt allemaal!

Leonid Padyukov, my main supervisor, for having faith in me as a student and for supporting me all these years. For allowing me the freedom to boast my own ideas and for guiding me on the path to independency by giving me the opportunity to contribute to the entire scientific process. I would also like to thank you for the advice you gave me to wear a bicycle helmet, even though I made it to the Christmas dinner without.

Vivianne Malmström, my co-supervisor, for the great scientific discussions and suggestions. You always made sure my science stayed on the right path towards T cells.

Karine Chemin, my co-supervisor, for your valuable support, comments and collaborations. You were always there to help me and it has truly been a pleasure to get to know you.

Anca Catrina, *Ingrid Lundberg* and *Lars Klareskog* for creating a nice environment at the Rheumatology Unit.

All current and former members of the Padyukov group – *Barbro Larsson, Espen Hesselberg*, *Gilad Silberberg*, *Klementy Shchetynsky*, *Lina Diaz-Gallo* and *Natalia Rivera* – I have learned a lot from all of you and it is safe to say that without you this time would not have been the same. *Barbro* for being the backbone of the group, *Espen* for the discussions about nothing and everything, *Gilad* for your valuable discussions, *Klementy* for literally being the roof above my head, *Lina* for being the opposite of me and of course the grilled peaches in the rain and *Natalia* for all your suggestions and for keeping me safe in Baltimore.

Boel Brynedal, *Daniel Ramsköld* and *Niyaz Yoosuf*, the non-official group members, for your support throughout the years.

Juan Galindo, *Kim Franson*, *Radovan Krejčíř*, *Sussan Alipour* and all other students that have passed by and I have had the chance to meet.

All of my co-authors – *Aase Hensvold*, *Anca Catrina*, *Ann Reed*, *Daniël Ramsköld*, *Espen Hesselberg*, *Ingrid Lundberg*, *Jessica Herrath*, *Karine Chemin*, *Karolina Tandre*, *Klementy Shchetynsky*, *Lars Klareskog*, *Lars Rönnblom*, *Leonid Padyukov*, *Lina Diaz-Gallo*, *Louise Ekholm*, *Maija-Leena Eloranta*, *Steffen Uebe* and *Vivianne Malmström* – for the pleasant collaborations and for guiding me through different disciplines of science.

Danika Schepis, *Louise Berg* and *Yvonne Sundström* for being there when I desperately needed antibodies and of course for all the other valuable help.

Gunnel Bemerfeldt, *Sanna Tavakoli*, *Stina Nordman* and *Veronica Ebbersten Lindholm* for their great help with all administrative issues.

The CMM IT service for helping out every time my computer or the network decided that it was not a good day for work.

Eva Jemseby, *Gloria Rostvall*, *Gull-Britt Almgren*, *Julia Norkko* and *Susana Hernandez Machado* not only for keeping track of all precious research samples, but also for entertaining me by performing experiments in my office.

Agnes Sowinska, *Bence Rethi*, *Meng Sun*, *Mia Olsson*, *Niyaz Yoosuf*, *Quan Tang* and *William Nyberg* for the warm and pleasant office environment and all other office rituals.

All other current and former members of the Rheumatology Unit, in particular *Adrian Levitsky*, *Akilan Krishnamurthy*, *Alina Johansson*, *Angeles Galindo*, *Cátia Cerqueira*, *Christina Gerstner*, *Guðný Ella Thorlacius*, *Heidi Wähämaa*, *Henna Salo*, *Joan Raouf*, *Johanna Estelius*, *Jorge Ramírez*, *Julia Steinmetz*, *Khaled Amara*, *Lara Mentlein*, *Marianne Engström*, *Natalia Sherina*, *Priya Revathikumar*, *Raya Saleh*, *Rita Ivanchenko*, *Sudeepta Kumar Panda*, *Uta Hardt* and *Vijay Joshua*, for the atmosphere and the conversations.

All the current and former members of the Neuroimmunology Unit, in particular *Eliane Piket*, *Lara Kular*, *Mohsen Khademi*, *Sabrina Ruhrmann* and *Xingmei Zhang*, for your contributions to the nice work environment.

Jan Alvar Lindencrona, *Helena Erlandsson Harris* and all the other members of the National Clinical Research School in Chronic Inflammatory Diseases (NCRSCID) for organization of and contributing to the educational and pleasant environment at all the meetings we have had together.

Björn Nystedt for accepting me in The Swedish Bioinformatics Advisory Program and my advisor *Jessica Lindvall* for the guidance in bioinformatics.

No words are needed to thank *Brian Timmer*, a mere second of eye contact is sufficient. My life would be incomplete without you!

7 REFERENCES

1. Smale ST, Kadonaga JT. The RNA polymerase II core promoter. Annu. Rev. Biochem., 2003;72:449.
2. Zabidi MA, Stark A. Regulatory Enhancer–Core-Promoter Communication via Transcription Factors and Cofactors. Trends Genet., 2016;32:801.
3. Whyte WA, Orlando DA, Hnisz D, Abraham BJ, Lin CY *et al.* Master Transcription Factors and Mediator Establish Super-Enhancers at Key Cell Identity Genes. Cell, 2013;153:307.
4. Vahedi G, Kanno Y, Furumoto Y, Jiang K, Parker SC *et al.* Stretch-Enhancers Delineate Disease-Associated Regulatory Nodes in T Cells. Nature, 2015;520:558.
5. Creyghton MP, Cheng AW, Welstead GG, Kooistra T, Carey BW *et al.* Histone H3K27ac separates active from poised enhancers and predicts developmental state. Proc. Natl. Acad. Sci. U.S.A., 2010;107:21931.
6. Spitz F, Furlong EEM. Transcription factors: from enhancer binding to developmental control. Nat. Rev. Genet., 2012;13:613.
7. Choi I, Kim R, Lim H-W, Kaestner KH, Won K-J. 5-hydroxymethylcytosine represses the activity of enhancers in embryonic stem cells: a new epigenetic signature for gene regulation. BMC Genomics, 2014;15:1.
8. Lomvardas S, Barnea G, Pisapia DJ, Mendelsohn M, Kirkland J *et al.* Interchromosomal Interactions and Olfactory Receptor Choice. Cell, 2006;126:403.
9. Arzate-Mejía RG, Recillas-Targa F, Corces VG. Developing in 3D: the role of CTCF in cell differentiation. Development, 2018;145.
10. Dixon JR, Selvaraj S, Yue F, Kim A, Li Y *et al.* Topological Domains in Mammalian Genomes Identified by Analysis of Chromatin Interactions. Nature, 2012;485:376.
11. Guo Y, Xu Q, Canzio D, Shou J, Li J *et al.* CRISPR Inversion of CTCF Sites Alters Genome Topology and Enhancer/Promoter Function. Cell, 2015;162:900.
12. Hu J, Zhang Y, Zhao L, Frock RL, Du Z *et al.* Chromosomal Loop Domains Direct the Recombination of Antigen Receptor Genes. Cell, 2015;163:947.
13. Shukla S, Kavak E, Gregory M, Imashimizu M, Shutinoski B *et al.* CTCF-promoted RNA polymerase II pausing links DNA methylation to splicing. Nature, 2011;479:74.
14. Juraeva D, Haenisch B, Zapatka M, Frank J, GROUP Investigators *et al.* Integrated Pathway-Based Approach Identifies Association between Genomic Regions at CTCF and CACNB2 and Schizophrenia. PLoS Genet., 2014;10:e1004345.
15. Kernohan KD, Vernimmen D, Gloor GB, Bérubé NG. Analysis of neonatal brain lacking ATRX or MeCP2 reveals changes in nucleosome density, CTCF binding and chromatin looping. Nucleic Acids Res., 2014;42:8356.

16. Kemp CJ, Moore JM, Moser R, Bernard B, Teater M *et al.* CTCF haploinsufficiency destabilizes DNA methylation and predisposes to cancer. Cell Rep., 2014;7:1020.
17. Wahl MC, Will CL, Lührmann R. The Spliceosome: Design Principles of a Dynamic RNP Machine. Cell, 2009;136:701.
18. Pan Q, Shai O, Lee LJ, Frey BJ, Blencowe BJ. Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. Nat. Genet., 2008;40:1413.
19. Wang ET, Sandberg R, Luo S, Khrebtukova I, Zhang L *et al.* Alternative isoform regulation in human tissue transcriptomes. Nature, 2008;456:470.
20. Baralle FE, Giudice J. Alternative splicing as a regulator of development and tissue identity. Nat. Rev. Mol. Cell Biol., 2017;18:437.
21. Amit M, Donyo M, Hollander D, Goren A, Kim E *et al.* Differential GC Content between Exons and Introns Establishes Distinct Strategies of Splice-Site Recognition. Cell Rep., 2012;1:543.
22. Schwartz S, Meshorer E, Ast G. Chromatin organization marks exon-intron structure. Nat. Struct. Mol. Biol., 2009;16:990.
23. Saint-André V, Batsché E, Rachez C, Muchardt C. Histone H3 lysine 9 trimethylation and HP1γ favor inclusion of alternative exons. Nat. Struct. Mol. Biol., 2011;18:337.
24. Huff JT, Plocik AM, Guthrie C, Yamamoto KR. Reciprocal intronic and exonic histone modification regions in humans. Nat. Struct. Mol. Biol., 2010;17:1495.
25. Trowbridge IS, Thomas ML. CD45: An Emerging Role as a Protein Tyrosine Phosphatase Required for Lymphocyte Activation and Development. Annu. Rev. Immunol., 1994;12:85.
26. Lejeune F, Li X, Maquat LE. Nonsense-Mediated mRNA Decay in Mammalian Cells Involves Decapping, Deadenylating, and Exonucleolytic Activities. Mol. Cell, 2003;12:675.
27. Mendell JT, Sharifi NA, Meyers JL, Martinez-Murillo F, Dietz HC. Nonsense surveillance regulates expression of diverse classes of mammalian transcripts and mutes genomic noise. Nat. Genet., 2004;36:1073.
28. Palade GE. A small particulate component of the cytoplasm. J. Biophys. Biochem. Cytol., 1955;1:59.
29. Hoagland MB, Stephenson ML, Scott JF, Hecht LI, Zamecnik PC. A soluble ribonucleic acid intermediate in protein synthesis. J. Biol. Chem., 1958;231:241.
30. International Human Genome Sequencing Consortium, Lander ES, Linton LM, Birren B, Nusbaum C *et al.* Initial sequencing and analysis of the human genome. Nature, 2001;409:860.
31. Venter JC, Adams MD, Myers EW, Li PW, Mural RJ *et al.* The Sequence of the Human Genome. Science, 2001;291:1304.
32. The Encode Project Consortium, Birney E, Stamatoyannopoulos JA, Dutta A, Guigó R *et al.* Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. Nature, 2007;447:799.

33. The Encode Project Consortium, Dunham I, Kundaje A, Aldred SF, Collins PJ *et al.* An integrated encyclopedia of DNA elements in the human genome. Nature, 2012;489:57.
34. Brown CJ, Hendrich BD, Rupert JL, Lafrenière RG, Xing Y *et al.* The human *XIST* gene: Analysis of a 17 kb inactive X-specific RNA that contains conserved repeats and is highly localized within the nucleus. Cell, 1992;71:527.
35. Rinn JL, Kertesz M, Wang JK, Squazzo SL, Xu X *et al.* Functional Demarcation of Active and Silent Chromatin Domains in Human HOX Loci by Non-Coding RNAs. Cell, 2007;129:1311.
36. Di Leva G, Garofalo M, Croce CM. MicroRNAs in Cancer. Annu. Rev. Pathol., 2014;9:287.
37. Sun E, Shi Y. MicroRNAs: Small molecules with big roles in neurodevelopment and diseases. Exp. Neurol., 2015;268:46.
38. Chen J-Q, Papp G, Szodoray P, Zeher M. The role of microRNAs in the pathogenesis of autoimmune diseases. Autoimmun. Rev., 2016;15:1171.
39. Wu G-C, Pan H-F, Leng R-X, Wang D-G, Li X-P *et al.* Emerging role of long noncoding RNAs in autoimmune diseases. Autoimmun. Rev., 2015;14:798.
40. Huarte M. The emerging role of lncRNAs in cancer. Nat. Med., 2015;21:1253.
41. Lee RC, Feinbaum RL, Ambros V. The C. elegans heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14. Cell, 1993;75:843.
42. Olsen PH, Ambros V. The lin-4 Regulatory RNA Controls Developmental Timing in Caenorhabditis elegans by Blocking LIN-14 Protein Synthesis after the Initiation of Translation. Dev. Biol., 1999;216:671.
43. Reinhart BJ, Slack FJ, Basson M, Pasquinelli AE, Bettinger JC *et al.* The 21-nucleotide let-7 RNA regulates developmental timing in Caenorhabditis elegans. Nature, 2000;403:901.
44. Lagos-Quintana M, Rauhut R, Yalcin A, Meyer J, Lendeckel W *et al.* Identification of Tissue-Specific MicroRNAs from Mouse. Curr. Biol., 2002;12:735.
45. Pasquinelli AE, Reinhart BJ, Slack F, Martindale MQ, Kuroda MI *et al.* Conservation of the sequence and temporal expression of let-7 heterochronic regulatory RNA. Nature, 2000;408:86.
46. Friedman RC, Farh KK-H, Burge CB, Bartel DP. Most mammalian mRNAs are conserved targets of microRNAs. Genome Res., 2009;19:92.
47. Shukla GC, Singh J, Barik S. MicroRNAs: Processing, Maturation, Target Recognition and Regulatory Functions. Mol. Cell. Pharmacol., 2011;3:83.
48. Guil S, Cáceres JF. The multifunctional RNA-binding protein hnRNP A1 is required for processing of miR-18a. Nat. Struct. Mol. Biol., 2007;14:591.
49. Elbashir SM, Lendeckel W, Tuschl T. RNA interference is mediated by 21- and 22-nucleotide RNAs. Genes Dev., 2001;15:188.
50. Sienski G, Dönertas D, Brennecke J. Transcriptional Silencing of Transposons by Piwi and Maelstrom and Its Impact on Chromatin State and Gene Expression. Cell, 2012;151:964.

51. Lin H, Spradling AC. A novel group of pumilio mutations affects the asymmetric division of germline stem cells in the Drosophila ovary. Development, 1997;124:2463.
52. Nandi S, Chandramohan D, Fioriti L, Melnick AM, Hébert JM *et al.* Roles for small noncoding RNAs in silencing of retrotransposons in the mammalian brain. Proc. Natl. Acad. Sci. U.S.A., 2016;113:12697.
53. Freedman JE, Gerstein M, Mick E, Rozowsky J, Levy D *et al.* Diverse human extracellular RNAs are widely detected in human plasma. Nat. Commun., 2016;7:11106.
54. DeChiara TM, Robertson EJ, Efstratiadis A. Parental imprinting of the mouse insulin-like growth factor II gene. Cell, 1991;64:849.
55. Bartolomei MS, Zemel S, Tilghman SM. Parental imprinting of the mouse H19 gene. Nature, 1991;351:153.
56. Brannan CI, Dees EC, Ingram RS, Tilghman SM. The product of the H19 gene may function as an RNA. Mol. Cell. Biol., 1990;10:28.
57. Lee J, Davidow LS, Warshawsky D. Tsix, a gene antisense to Xist at the X-inactivation centre. Nat. Genet., 1999;21:400.
58. Kapranov P, Cawley SE, Drenkow J, Bekiranov S, Strausberg RL *et al.* Large-Scale Transcriptional Activity in Chromosomes 21 and 22. Science, 2002;296:916.
59. Rinn JL, Euskirchen G, Bertone P, Martone R, Luscombe NM *et al.* The transcriptional activity of human Chromosome 22. Genes Dev., 2003;17:529.
60. The Fantom Consortium, The Riken Genome Exploration Research Group Phase I & II Team, Okazaki Y, Furuno M, Kasukawa T *et al.* Analysis of the mouse transcriptome based on functional annotation of 60,770 full-length cDNAs. Nature, 2002;420:563.
61. Katayama S, Tomaru Y, Kasukawa T, Waki K, Nakanishi M *et al.* Antisense Transcription in the Mammalian Transcriptome. Science, 2005;309:1564.
62. Shiraki T, Kondo S, Katayama S, Waki K, Kasukawa T *et al.* Cap analysis gene expression for high-throughput analysis of transcriptional starting point and identification of promoter usage. Proc. Natl. Acad. Sci. U.S.A., 2003;100:15776.
63. Kawaji H, Kasukawa T, Forrest A, Carninci P, Hayashizaki Y. The FANTOM5 collection, a data series underpinning mammalian transcriptome atlases in diverse cell types. Sci. Data, 2017;4:170113.
64. Hon C-C, Ramilowski JA, Harshbarger J, Bertin N, Rackham OJL *et al.* An atlas of human long non-coding RNAs with accurate 5′ ends. Nature, 2017;543:199.
65. Djebali S, Davis CA, Merkel A, Dobin A, Lassmann T *et al.* Landscape of transcription in human cells. Nature, 2012;489:101.
66. Bánfai B, Jia H, Khatun J, Wood E, Risk B *et al.* Long noncoding RNAs are rarely translated in two human cell lines. Genome Res., 2012;22:1646.
67. Derrien T, Johnson R, Bussotti G, Tanzer A, Djebali S *et al.* The GENCODE v7 catalog of human long noncoding RNAs: Analysis of their gene structure, evolution, and expression. Genome Res., 2012;22:1775.

68. Tsai M-C, Manor O, Wan Y, Mosammaparast N, Wang JK *et al.* Long Noncoding RNA as Modular Scaffold of Histone Modification Complexes. Science, 2010;329:689.
69. Salmena L, Poliseno L, Tay Y, Kats L, Pandolfi PP. A ceRNA hypothesis: the Rosetta stone of a hidden RNA language? Cell, 2011;146:353.
70. Ebert MS, Neilson JR, Sharp PA. MicroRNA sponges: competitive inhibitors of small RNAs in mammalian cells. Nat. Methods, 2007;4:721.
71. Poliseno L, Salmena L, Zhang J, Carver B, Haveman WJ *et al.* A coding-independent function of gene and pseudogene mRNAs regulates tumour biology. Nature, 2010;465:1033.
72. Plath K, Fang J, Mlynarczyk-Evans SK, Cao R, Worringer KA *et al.* Role of Histone H3 Lysine 27 Methylation in X Inactivation. Science, 2003;300:131.
73. Sun BK, Deaton AM, Lee JT. A Transient Heterochromatic State in Xist Preempts X Inactivation Choice without RNA Stabilization. Mol. Cell, 2006;21:617.
74. Feng J, Bi C, Clark BS, Mady R, Shah P *et al.* The Evf-2 noncoding RNA is transcribed from the Dlx-5/6 ultraconserved region and functions as a Dlx-2 transcriptional coactivator. Genes Dev., 2006;20:1470.
75. Berghoff EG, Clark MF, Chen S, Cajigas I, Leib DE *et al.* Evf2 (Dlx6as) lncRNA regulates ultraconserved enhancer methylation and the differential transcriptional control of adjacent genes. Development, 2013;140:4407.
76. Marques AC, Hughes J, Graham B, Kowalczyk MS, Higgs DR *et al.* Chromatin signatures at transcriptional start sites separate two equally populated yet distinct classes of intergenic long noncoding RNAs. Genome Biol., 2013;14:R131.
77. Martianov I, Ramadass A, Serra Barros A, Chow N, Akoulitchev A. Repression of the human dihydrofolate reductase gene by a non-coding interfering transcript. Nature, 2007;445:666.
78. Tripathi V, Ellis JD, Shen Z, Song DY, Pan Q *et al.* The Nuclear-Retained Noncoding RNA MALAT1 Regulates Alternative Splicing by Modulating SR Splicing Factor Phosphorylation. Mol. Cell, 2010;39:925.
79. Beltran M, Puig I, Peña C, García JM, Álvarez AB *et al.* A natural antisense transcript regulates Zeb2/Sip1 gene expression during Snail1-induced epithelial–mesenchymal transition. Genes Dev., 2008;22:756.
80. Faghihi MA, Zhang M, Huang J, Modarresi F, Van der Brug MP *et al.* Evidence for natural antisense transcript-mediated inhibition of microRNA function. Genome Biol., 2010;11:R56.
81. Gong C, Maquat LE. lncRNAs transactivate Staufen1-mediated mRNA decay by duplexing with 3'UTRs via Alu elements. Nature, 2011;470:284.
82. Kim YK, Furic L, DesGroseillers L, Maquat LE. Mammalian Staufen1 Recruits Upf1 to Specific mRNA 3′UTRs so as to Elicit mRNA Decay. Cell, 2005;120:195.
83. The Genomes Project Consortium, Durbin RM, Altshuler D, Durbin RM, Abecasis GR *et al.* A map of human genome variation from population-scale sequencing. Nature, 2010;467:1061.

84. The Genomes Project Consortium, McVean GA, Altshuler DM, Durbin RM, Abecasis GR *et al.* An integrated map of genetic variation from 1,092 human genomes. Nature, 2012;491:56.
85. The International HapMap Consortium, Gibbs RA, Belmont JW, Hardenbol P, Willis TD *et al.* The International HapMap Project. Nature, 2003;426:789.
86. The International HapMap Consortium, Altshuler D, Donnelly P. A haplotype map of the human genome. Nature, 2005;437:1299.
87. The Genomes Project Consortium, Auton A, Abecasis GR, Altshuler DM, Durbin RM *et al.* A global reference for human genetic variation. Nature, 2015;526:68.
88. Kimchi-Sarfaty C, Oh JM, Kim I-W, Sauna ZE, Calcagno AM *et al.* A "Silent" Polymorphism in the *MDR*1 Gene Changes Substrate Specificity. Science, 2007;315:525.
89. Cargill M, Altshuler D, Ireland J, Sklar P, Ardlie K *et al.* Characterization of single-nucleotide polymorphisms in coding regions of human genes. Nat. Genet., 1999;22:231.
90. Coulombe-Huntington J, Lam KCL, Dias C, Majewski J. Fine-Scale Variation and Genetic Determinants of Alternative Splicing across Individuals. PLoS Genet., 2009;5:e1000766.
91. Zhernakova DV, Deelen P, Vermaat M, van Iterson M, van Galen M *et al.* Identification of context-dependent expression quantitative trait loci in whole blood. Nat. Genet., 2016;49:139.
92. Martin-Trujillo A, Iglesias-Platas I, Coto E, Corral-Juan M, San Nicolás H *et al.* Genotype of an individual single nucleotide polymorphism regulates DNA methylation at the TRPC3 alternative promoter. Epigenetics, 2011;6:1236.
93. Ozaki K, Ohnishi Y, Iida A, Sekine A, Yamada R *et al.* Functional SNPs in the lymphotoxin-α gene that are associated with susceptibility to myocardial infarction. Nat. Genet., 2002;32:650.
94. Haines JL, Hauser MA, Schmidt S, Scott WK, Olson LM *et al.* Complement Factor H Variant Increases the Risk of Age-Related Macular Degeneration. Science, 2005;308:419.
95. The Wellcome Trust Case Control Consortium. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature, 2007;447:661.
96. Welter D, MacArthur J, Morales J, Burdett T, Hall P *et al.* The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. Nucleic Acids Res., 2014;42:D1001.
97. MacArthur J, Bowler E, Cerezo M, Gil L, Hall P *et al.* The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog). Nucleic Acids Res., 2017;45:D896.
98. Cortes A, Brown MA. Promise and pitfalls of the Immunochip. Arthritis Res. Ther., 2011;13:101.
99. Voight BF, Kang HM, Ding J, Palmer CD, Sidore C *et al.* The Metabochip, a Custom Genotyping Array for Genetic Studies of Metabolic, Cardiovascular, and Anthropometric Traits. PLoS Genet., 2012;8:e1002793.

100. Grundberg E, Small KS, Hedman ÅK, Nica AC, Buil A *et al.* Mapping cis- and trans-regulatory effects across multiple tissues in twins. Nat. Genet., 2012;44:1084.
101. Wright FA, Sullivan PF, Brooks AI, Zou F, Sun W *et al.* Heritability and genomics of gene expression in peripheral blood. Nat. Genet., 2014;46:430.
102. Harismendy O, Notani D, Song X, Rahim NG, Tanasa B *et al.* 9p21 DNA variants associated with Coronary Artery Disease impair IFNγ signaling response. Nature, 2011;470:264.
103. Pomerantz MM, Ahmadiyeh N, Jia L, Herman P, Verzi MP *et al.* The 8q24 cancer risk variant rs6983267 shows long-range interaction with MYC in colorectal cancer. Nat. Genet., 2009;41:882.
104. Martin MP, Dean M, Smith MW, Winkler C, Gerrard B *et al.* Genetic Acceleration of AIDS Progression by a Promoter Variant of CCR5. Science, 1998;282:1907.
105. The Riken Genome Exploration Research Group Phase II Team, The Fantom Consortium, Kawai J, Shinagawa A, Shibata K *et al.* Functional annotation of a full-length mouse cDNA collection. Nature, 2001;409:685.
106. Bernstein BE, Stamatoyannopoulos JA, Costello JF, Ren B, Milosavljevic A *et al.* The NIH Roadmap Epigenomics Mapping Consortium. Nat. Biotechnol., 2010;28.
107. Lonsdale J, Thomas J, Salvatore M, Phillips R, Lo E *et al.* The Genotype-Tissue Expression (GTEx) project. Nat. Genet., 2013;45:580.
108. Nica AC, Montgomery SB, Dimas AS, Stranger BE, Beazley C *et al.* Candidate Causal Regulatory Effects by Integration of Expression QTLs with Complex Trait Genetic Associations. PLoS Genet., 2010;6:e1000895.
109. Nicolae DL, Gamazon E, Zhang W, Duan S, Dolan ME *et al.* Trait-Associated SNPs Are More Likely to Be eQTLs: Annotation to Enhance Discovery from GWAS. PLoS Genet., 2010;6:e1000888.
110. GTEx Consortium, Aguet F, Brown AA, Castel SE, Davis JR *et al.* Genetic effects on gene expression across human tissues. Nature, 2017;550:204.
111. Stranger BE, Montgomery SB, Dimas AS, Parts L, Stegle O *et al.* Patterns of Cis Regulatory Variation in Diverse Human Populations. PLoS Genet., 2012;8:e1002639.
112. Frayling TM, Timpson NJ, Weedon MN, Zeggini E, Freathy RM *et al.* A Common Variant in the FTO Gene Is Associated with Body Mass Index and Predisposes to Childhood and Adult Obesity. Science, 2007;316:889.
113. Scuteri A, Sanna S, Chen W-M, Uda M, Albai G *et al.* Genome-Wide Association Scan Shows Genetic Variants in the FTO Gene Are Associated with Obesity-Related Traits. PLoS Genet., 2007;3:e115.
114. Dina C, Meyre D, Gallina S, Durand E, Körner A *et al.* Variation in FTO contributes to childhood obesity and severe adult obesity. Nat. Genet., 2007;39:724.
115. Claussnitzer M, Dankel SN, Kim K-H, Quon G, Meuleman W *et al.* FTO Obesity Variant Circuitry and Adipocyte Browning in Humans. N. Engl. J. Med., 2015;373:895.

116. Lappalainen T, Sammeth M, Friedlander MR, ’t Hoen PAC, Monlong J *et al.* Transcriptome and genome sequencing uncovers functional variation in humans. Nature, 2013;501:506.
117. Heinig M, Petretto E, Wallace C, Bottolo L, Rotival M *et al.* A trans-acting locus regulates an anti-viral expression network and type 1 diabetes risk. Nature, 2010;467:460.
118. The MuTHER Consortium, Small KS, Hedman ÅK, Grundberg E, Nica AC *et al.* Identification of an imprinted master trans regulator at the KLF14 locus related to multiple metabolic phenotypes. Nat. Genet., 2011;43:561.
119. Westra H-J, Peters MJ, Esko T, Yaghootkar H, Schurmann C *et al.* Systematic identification of trans eQTLs as putative drivers of known disease associations. Nat. Genet., 2013;45:1238.
120. Fairfax BP, Makino S, Radhakrishnan J, Plant K, Leslie S *et al.* Genetics of gene expression in primary immune cells identifies cell type–specific master regulators and roles of HLA alleles. Nat. Genet., 2012;44:502.
121. Fu J, Wolfs MGM, Deelen P, Westra H-J, Fehrmann RSN *et al.* Unraveling the Regulatory Mechanisms Underlying Tissue-Dependent Genetic Variation of Gene Expression. PLoS Genet., 2012;8:e1002431.
122. Brown CD, Mangravite LM, Engelhardt BE. Integrative Modeling of eQTLs and Cis-Regulatory Elements Suggests Mechanisms Underlying Cell Type Specificity of eQTLs. PLoS Genet., 2013;9:e1003649.
123. Fairfax BP, Humburg P, Makino S, Naranbhai V, Wong D *et al.* Innate Immune Activity Conditions the Effect of Regulatory Variants upon Monocyte Gene Expression. Science, 2014;343:1246949.
124. Ferraro A, D’Alise AM, Raj T, Asinovski N, Phillips R *et al.* Interindividual variation in human T regulatory cells. Proc. Natl. Acad. Sci. U.S.A., 2014;111:E1111.
125. Lee MN, Ye C, Villani A-C, Raj T, Li W *et al.* Common Genetic Variants Modulate Pathogen-Sensing Responses in Human Dendritic Cells. Science, 2014;343:1246980.
126. Battle A, Mostafavi S, Zhu X, Potash JB, Weissman MM *et al.* Characterizing the genetic basis of transcriptome diversity through RNA-sequencing of 922 individuals. Genome Res., 2014;24:14.
127. Takata A, Matsumoto N, Kato T. Genome-wide identification of splicing QTLs in the human brain and their enrichment among schizophrenia-associated loci. Nat. Commun., 2017;8:14519.
128. Fadista J, Vikman P, Laakso EO, Mollet IG, Esguerra JL *et al.* Global genomic and transcriptomic analysis of human pancreatic islets reveals novel genes influencing glucose metabolism. Proc. Natl. Acad. Sci. U.S.A., 2014;111:13924.
129. Franke A, McGovern DPB, Barrett JC, Wang K, Radford-Smith GL *et al.* Genome-wide meta-analysis increases to 71 the number of confirmed Crohn’s disease susceptibility loci. Nat. Genet., 2010;42:1118.
130. Suhre K, Arnold M, Bhagwat AM, Cotton RJ, Engelke R *et al.* Connecting genetic risk to disease end points through the human blood plasma proteome. Nat. Commun., 2017;8:14357.

131. Sun BB, Maranville JC, Peters JE, Stacey D, Staley JR *et al.* Genomic atlas of the human plasma proteome. Nature, 2018;558:73.
132. Battle A, Khan Z, Wang SH, Mitrano A, Ford MJ *et al.* Impact of Regulatory Variation from RNA to Protein. Science, 2015;347:664.
133. Cenik C, Sarinay Cenik E, Byeon GW, Grubert F, Candille SI *et al.* Integrative analysis of RNA, translation and protein levels reveals distinct regulatory variation across humans. Genome Res., 2015;25:1610.
134. Degner JF, Pai AA, Pique-Regi R, Veyrieras J-B, Gaffney DJ *et al.* DNaseI sensitivity QTLs are a major determinant of human expression variation. Nature, 2012;482:390.
135. McVicker G, van de Geijn B, Degner JF, Cain CE, Banovich NE *et al.* Identification of Genetic Variants That Affect Histone Modifications in Human Cells. Science, 2013;342:747.
136. Bell JT, Pai AA, Pickrell JK, Gaffney DJ, Pique-Regi R *et al.* DNA methylation patterns associate with genetic and gene expression variation in HapMap cell lines. Genome Biol., 2011;12:R10.
137. Gaunt TR, Shihab HA, Hemani G, Min JL, Woodward G *et al.* Systematic identification of genetic influences on methylation across the human life course. Genome Biol., 2016;17:61.
138. Do C, Lang Charles F, Lin J, Darbary H, Krupska I *et al.* Mechanisms and Disease Associations of Haplotype-Dependent Allele-Specific DNA Methylation. Am. J. Hum. Genet., 2016;98:934.
139. Wu Y, Zeng J, Zhang F, Zhu Z, Qi T *et al.* Integrative analysis of omics summary data reveals putative mechanisms underlying complex traits. Nat. Commun., 2018;9:918.
140. Abbas AK. Basic immunology functions and disorders of the immune system, 4th edn. Philadelphia: Elsevier/Saunders; 2014.
141. Kawai T, Akira S. The roles of TLRs, RLRs and NLRs in pathogen recognition. Int. Immunol., 2009;21:317.
142. Ricklin D, Hajishengallis G, Yang K, Lambris JD. Complement: a key system for immune surveillance and homeostasis. Nat. Immunol., 2010;11:785.
143. Asselin-Paturel C, Trinchieri G. Production of type I interferons: plasmacytoid dendritic cells and beyond. J. Exp. Med., 2005;202:461.
144. Topham NJ, Hewitt EW. Natural killer cell cytotoxicity: how do they pull the trigger? Immunology, 2009;128:7.
145. Rock KL, Shen L. Cross-presentation: underlying mechanisms and role in immune surveillance. Immunol Rev., 2005;207:166.
146. Chen L, Flies DB. Molecular mechanisms of T cell co-stimulation and co-inhibition. Nat. Rev. Immunol., 2013;13:227.
147. Luckheeram RV, Zhou R, Verma AD, Xia B. $CD4^{+}$T cells: differentiation and functions. Clin. Dev. Immunol., 2012;2012:925135.
148. Aune TM, Collins PL, Chang S. Epigenetics and T helper 1 differentiation. Immunology, 2009;126:299.
149. Crotty S. Follicular Helper CD4 T Cells (TFH). Annu. Rev. Immunol., 2011;29:621.
150. Rudensky AY. Regulatory T cells and Foxp3. Immunol Rev., 2011;241:260.

151. Zhang N, Bevan MJ. CD8(+) T cells: foot soldiers of the immune system. Immunity, 2011;35:161.
152. Kyewski B, Klein L. A central role for central tolerance. Annu. Rev. Immunol., 2006;24:571.
153. Anderson MS, Venanzi ES, Klein L, Chen Z, Berzins SP *et al.* Projection of an Immunological Self Shadow Within the Thymus by the Aire Protein. Science, 2002;298:1395.
154. Sakaguchi S, Yamaguchi T, Nomura T, Ono M. Regulatory T Cells and Immune Tolerance. Cell, 2008;133:775.
155. Rosenblum MD, Remedios KA, Abbas AK. Mechanisms of human autoimmunity. J. Clin. Invest., 2015;125:2228.
156. Ziff M. Relation of cellular infiltration of rheumatoid synovial membrane to its immune response. Arthritis Rheum., 1974;17:313.
157. Firestein GS. Evolving concepts of rheumatoid arthritis. Nature, 2003;423:356.
158. Rose HM, Ragan C, Pearce E, Lipman MO. Differential Agglutination of Normal and Sensitized Sheep Erythrocytes by Sera of Patients with Rheumatoid Arthritis. Proc. Soc. Exp. Biol. Med., 1948;68:1.
159. Schellekens GA, de Jong BA, van den Hoogen FH, van de Putte LB, van Venrooij WJ. Citrulline is an essential constituent of antigenic determinants recognized by rheumatoid arthritis-specific autoantibodies. J. Clin. Invest., 1998;101:273.
160. Prete M, Racanelli V, Digiglio L, Vacca A, Dammacco F *et al.* Extra-articular manifestations of rheumatoid arthritis: An update. Autoimmun. Rev., 2011;11:123.
161. Scott DL, Wolfe F, Huizinga TWJ. Rheumatoid arthritis. Lancet, 2010;376:1094.
162. Smolen JS, Landewé R, Bijlsma J, Burmester G, Chatzidionysiou K *et al.* EULAR recommendations for the management of rheumatoid arthritis with synthetic and biological disease-modifying antirheumatic drugs: 2016 update. Ann. Rheum. Dis., 2017;76:960.
163. Aletaha D, Neogi T, Silman AJ, Funovits J, Felson DT *et al.* 2010 Rheumatoid arthritis classification criteria: an American College of Rheumatology/European League Against Rheumatism collaborative initiative. Ann. Rheum. Dis., 2010;69:1580.
164. Silman AJ, Macgregor AJ, Thomson W, Holligan S, Carthy D *et al.* Twin concordance rates for rheumatoid arthritis: results from a nationwide study. Rheumatology, 1993;32:903.
165. Terao C, Ikari K, Nakayamada S, Takahashi Y, Yamada R *et al.* A twin study of rheumatoid arthritis in the Japanese population. Mod. Rheumatol., 2016;26:685.
166. Svendsen AJ, Kyvik KO, Houen G, Junker P, Christensen K *et al.* On the Origin of Rheumatoid Arthritis: The Impact of Environment and Genes - A Population Based Twin Study. PLoS One, 2013;8:e57304.
167. Okada Y, Wu D, Trynka G, Raj T, Terao C *et al.* Genetics of rheumatoid arthritis contributes to biology and drug discovery. Nature, 2014;506:376.

168. Cobb JE, Plant D, Flynn E, Tadjeddine M, Dieudé P *et al.* Identification of the Tyrosine-Protein Phosphatase Non-Receptor Type 2 as a Rheumatoid Arthritis Susceptibility Locus in Europeans. PLoS One, 2013;8:e66456.
169. Raychaudhuri S, Thomson BP, Remmers EF, Eyre S, Hinks A *et al.* Genetic variants at CD28, PRDM1 and CD2/CD58 are associated with rheumatoid arthritis risk. Nat. Genet., 2009;41:1313.
170. Gregersen PK, Silver J, Winchester RJ. The shared epitope hypothesis. an approach to understanding the molecular genetics of susceptibility to rheumatoid arthritis. Arthritis Rheum., 1987;30:1205.
171. Raychaudhuri S, Sandor C, Stahl EA, Freudenberg J, Lee H-S *et al.* Five amino acids in three HLA proteins explain most of the association between MHC and seropositive rheumatoid arthritis. Nat. Genet., 2012;44:291.
172. Scally SW, Petersen J, Law SC, Dudek NL, Nel HJ *et al.* A molecular basis for the association of the HLA-DRB1 locus, citrullination, and rheumatoid arthritis. J. Exp. Med., 2013;210:2569.
173. Gonzalez-Gay MA, Garcia-Porrua C, Hajeer AH. Influence of human leukocyte antigen-DRB1 on the susceptibility and severity of rheumatoid arthritis. Semin. Arthritis Rheum., 2002;31:355.
174. Weyand CM, Hicok KC, Conn DL, Goronzy JJ. THe influence of hla-drb1 genes on disease severity in rheumatoid arthritis. Ann. Intern. Med., 1992;117:801.
175. Mewar D, Marinou I, Coote AL, Moore DJ, Akil M *et al.* Association between radiographic severity of rheumatoid arthritis and shared epitope alleles: differing mechanisms of susceptibility and protection. Ann. Rheum. Dis., 2008;67:980.
176. van der Woude D, Lie BA, Lundström E, Balsa A, Feitsma AL *et al.* Protection against anti–citrullinated protein antibody–positive rheumatoid arthritis is predominantly associated with HLA–DRB1*1301: A meta-analysis of HLA–DRB1 associations with anti–citrullinated protein antibody–positive and anti–citrullinated protein antibody–negative rheumatoid arthritis in four European populations. Arthritis Rheum., 2010;62:1236.
177. Lundström E, Källberg H, Smolnikova M, Ding B, Rönnelid J *et al.* Opposing effects of HLA–DRB1*13 alleles on the risk of developing anti–citrullinated protein antibody–positive and anti–citrullinated protein antibody–negative rheumatoid arthritis. Arthritis Rheum., 2009;60:924.
178. Källberg H, Padyukov L, Plenge RM, Rönnelid J, Gregersen PK *et al.* Gene-Gene and Gene-Environment Interactions Involving HLA-DRB1, PTPN22, and Smoking in Two Subsets of Rheumatoid Arthritis. Am. J. Hum. Genet., 2007;80:867.
179. Diaz-Gallo L-M, Ramsköld D, Shchetynsky K, Folkersen L, Chemin K *et al.* Systematic approach demonstrates enrichment of multiple interactions between non-HLA risk variants and HLA-DRB1 risk alleles in rheumatoid arthritis. Ann. Rheum. Dis., 2018;77:1454.
180. Shchetynsky K, Protsyuk D, Ronninger M, Diaz-Gallo L-M, Klareskog L *et al.* Gene–gene interaction and RNA splicing profiles of MAP2K4 gene in rheumatoid arthritis. Clin. Immunol., 2015;158:19.

181. Bottini N, Musumeci L, Alonso A, Rahmouni S, Nika K *et al.* A functional variant of lymphoid tyrosine phosphatase is associated with type I diabetes. Nat. Genet., 2004;36:337.
182. Criswell LA, Pfeiffer KA, Lum RF, Gonzales B, Novitzke J *et al.* Analysis of Families in the Multiple Autoimmune Disease Genetics Consortium (MADGC) Collection: the PTPN22 620W Allele Associates with Multiple Autoimmune Phenotypes. Am. J. Hum. Genet., 2005;76:561.
183. Kyogoku C, Langefeld CD, Ortmann WA, Lee A, Selby S *et al.* Genetic Association of the R620W Polymorphism of Protein Tyrosine Phosphatase PTPN22 with Human SLE. Am. J. Hum. Genet., 2004;75:504.
184. Rothwell S, Cooper RG, Lundberg IE, Miller FW, Gregersen PK *et al.* Dense Genotyping of Immune-Related Loci in the Idiopathic Inflammatory Myopathies Confirms HLA alleles as Strongest Genetic Risk Factor and Suggests Different Genetic Background for Major Clinical Subgroups. Ann. Rheum. Dis., 2016;75:1558.
185. Begovich AB, Carlton VEH, Honigberg LA, Schrodi SJ, Chokkalingam AP *et al.* A Missense Single-Nucleotide Polymorphism in a Gene Encoding a Protein Tyrosine Phosphatase (PTPN22) Is Associated with Rheumatoid Arthritis. Am. J. Hum. Genet., 2004;75:330.
186. Plenge RM, Padyukov L, Remmers EF, Purcell S, Lee AT *et al.* Replication of Putative Candidate-Gene Associations with Rheumatoid Arthritis in >4,000 Samples from North America and Sweden: Association of Susceptibility with PTPN22, CTLA4, and PADI4. Am. J. Hum. Genet., 2005;77:1044.
187. Begovich AB, Caillier SJ, Alexander HC, Penko JM, Hauser SL *et al.* The R620W Polymorphism of the Protein Tyrosine Phosphatase PTPN22 Is Not Associated with Multiple Sclerosis. Am. J. Hum. Genet., 2005;76:184.
188. Diaz-Gallo L-M, Espino-Paisán L, Fransen K, Gómez-García M, van Sommeren S *et al.* Differential association of two PTPN22 coding variants with Crohn's disease and ulcerative colitis. Inflamm. Bowel Dis., 2011;17:2287.
189. Cohen S, Dadi H, Shaoul E, Sharfe N, Roifman CM. Cloning and Characterization of a Lymphoid-Specific, Inducible Human Protein Tyrosine Phosphatase, Lyp. Blood, 1999;93:2013.
190. de la Puerta ML, Trinidad AG, Rodríguez MdC, de Pereda JM, Sánchez Crespo M *et al.* The Autoimmunity Risk Variant LYP-W620 Cooperates with CSK in the Regulation of TCR Signaling. PLoS One, 2013;8:e54569.
191. Wu J, Katrekar A, Honigberg LA, Smith AM, Conn MT *et al.* Identification of Substrates of Human Protein-tyrosine Phosphatase PTPN22. J. Biol. Chem., 2006;281:11002.
192. Hill RJ, Zozulya S, Lu Y-L, Ward K, Gishizky M *et al.* The lymphoid protein tyrosine phosphatase Lyp interacts with the adaptor molecule Grb2 and functions as a negative regulator of T-cell activation. Exp. Hematol., 2002;30:237.
193. Rieck M, Arechiga A, Onengut-Gumuscu S, Greenbaum C, Concannon P *et al.* Genetic Variation in PTPN22 Corresponds to Altered Function of T and B Lymphocytes. J. Immunol., 2007;179:4704.

194. Zikherman J, Hermiston M, Steiner D, Hasegawa K, Chan A *et al.* PTPN22 deficiency cooperates with the CD45 E613R allele to break tolerance on a non-autoimmune background. J. Immunol., 2009;182:4093.
195. Zhang J, Zahir N, Jiang Q, Miliotis H, Heyraud S *et al.* The autoimmune disease–associated PTPN22 variant promotes calpain-mediated Lyp/Pep degradation associated with lymphocyte and dendritic cell hyperresponsiveness. Nat. Genet., 2011;43:902.
196. Dai X, James RG, Habib T, Singh S, Jackson S *et al.* A disease-associated PTPN22 variant promotes systemic autoimmunity in murine models. J. Clin. Invest., 2013;123:2024.
197. Wang Y, Shaked I, Stanford SM, Zhou W, Curtsinger JM *et al.* The Autoimmunity-Associated Gene PTPN22 Potentiates Toll-like Receptor-Driven, Type 1 Interferon-Dependent Immunity. Immunity, 2013;39:111.
198. Todd JA, Walker NM, Cooper JD, Smyth DJ, Downes K *et al.* Robust associations of four new chromosome regions from genome-wide analyses of type 1 diabetes. Nat. Genet., 2007;39:857.
199. Smyth DJ, Plagnol V, Walker NM, Cooper JD, Downes K *et al.* Shared and Distinct Genetic Variants in Type 1 Diabetes and Celiac Disease. N. Engl. J. Med., 2008;359:2767.
200. Thompson SD, Sudman M, Ramos PS, Marion MC, Ryan M *et al.* The susceptibility loci Juvenile Idiopathic Arthritis shares with other autoimmune diseases extend to PTPN2, COG6 and ANGPT1. Arthritis Rheum., 2010;62:3265.
201. Zhernakova A, Stahl EA, Trynka G, Raychaudhuri S, Festen EA *et al.* Meta-Analysis of Genome-Wide Association Studies in Celiac Disease and Rheumatoid Arthritis Identifies Fourteen Non-HLA Shared Loci. PLoS Genet., 2011;7:e1002004.
202. Eyre S, Bowes J, Diogo D, Lee A, Barton A *et al.* High density genetic mapping identifies new susceptibility loci for rheumatoid arthritis. Nat. Genet., 2012;44:1336.
203. Freudenberg J, Lee H-S, Han B-G, Shin HD, Kang YM *et al.* Genome-wide association study of rheumatoid arthritis in Koreans: Population-specific loci as well as overlap with European susceptibility loci. Arthritis Rheum., 2011;63:884.
204. Heinonen KM, Nestel FP, Newell EW, Charette G, Seemayer TA *et al.* T-cell protein tyrosine phosphatase deletion results in progressive systemic inflammatory disease. Blood, 2004;103:3457.
205. Simoncic PD, Lee-Loy A, Barber DL, Tremblay ML, McGlade CJ. The T Cell Protein Tyrosine Phosphatase Is a Negative Regulator of Janus Family Kinases 1 and 3. Curr. Biol., 2002;12:446.
206. ten Hoeve J, de Jesus Ibarra-Sanchez M, Fu Y, Zhu W, Tremblay M *et al.* Identification of a Nuclear Stat1 Protein Tyrosine Phosphatase. Mol. Cell. Biol., 2002;22:5662.
207. Long SA, Cerosaletti K, Wan JY, Ho JC, Tatum M *et al.* An autoimmune-associated variant in PTPN2 reveals an impairment of IL-2R signaling in CD4+ T cells. Genes Immun., 2011;12:116.

208. Böhmer F-D, Friedrich K. Protein tyrosine phosphatases as wardens of STAT signaling. JAKSTAT, 2014;3:e28087.
209. Wiede F, Shields BJ, Chew SH, Kyparissoudis K, van Vliet C *et al.* T cell protein tyrosine phosphatase attenuates T cell signaling to maintain tolerance in mice. J. Clin. Invest., 2011;121:4758.
210. Di Giuseppe D, Discacciati A, Orsini N, Wolk A. Cigarette smoking and risk of rheumatoid arthritis: a dose-response meta-analysis. Arthritis Res. Ther., 2014;16:R61.
211. Stolt P, Bengtsson C, Nordmark B, Lindblad S, Lundberg I *et al.* Quantification of the influence of cigarette smoking on rheumatoid arthritis: results from a population based case-control study, using incident cases. Ann. Rheum. Dis., 2003;62:835.
212. Karlson EW, Lee IM, Cook NR, Manson JE, Buring JE *et al.* A retrospective cohort study of cigarette smoking and risk of rheumatoid arthritis in female health professionals. Arthritis Rheum., 2001;42:910.
213. Källberg H, Ding B, Padyukov L, Bengtsson C, Rönnelid J *et al.* Smoking is a major preventable risk factor for Rheumatoid arthritis Estimations of risks after various exposures to cigarette smoke. Ann. Rheum. Dis., 2011;70:508.
214. Lundström E, Källberg H, Alfredsson L, Klareskog L, Padyukov L. Gene-environment interaction between DRB1 shared epitope and smoking regarding risk of ACPA-positive rheumatoid arthritis - all alleles are important. Arthritis Rheum., 2009;60:1597.
215. Costenbader KH, Chang S-C, De Vivo I, Plenge R, Karlson EW. Genetic polymorphisms in PTPN22, PADI-4, and CTLA-4 and risk for rheumatoid arthritis in two longitudinal cohort studies: evidence of gene-environment interactions with heavy cigarette smoking. Arthritis Res. Ther., 2008;10:R52.
216. Stolt P, Källberg H, Lundberg I, Sjögren B, Klareskog L *et al.* Silica exposure is associated with increased risk of developing rheumatoid arthritis: results from the Swedish EIRA study. Ann. Rheum. Dis., 2005;64:582.
217. Stolt P, Yahya A, Bengtsson C, Källberg H, Rönnelid J *et al.* Silica exposure among male current smokers is associated with a high risk of developing ACPA-positive rheumatoid arthritis. Ann. Rheum. Dis., 2010;69:1072.
218. Sverdrup B, Källberg H, Bengtsson C, Lundberg I, Padyukov L *et al.* Association between occupational exposure to mineral oil and rheumatoid arthritis: results from the Swedish EIRA case–control study. Arthritis Res. Ther., 2005;7:R1296.
219. Too CL, Muhamad NA, Ilar A, Padyukov L, Alfredsson L *et al.* Occupational exposure to textile dust increases the risk of rheumatoid arthritis: results from a Malaysian population-based case–control study. Ann. Rheum. Dis., 2016;75:997.
220. Caplan L, Davis LA, Bright CM, Kerr GS, Lazaro DM *et al.* Body mass index and the rheumatoid arthritis swollen joint count: an observational study. Arthritis Care Res., 2013;65:101.
221. Lu B, Hiraki L, Sparks JA, Malspeis S, Chen C-Y *et al.* Being overweight or obese and risk of developing rheumatoid arthritis among women: a prospective cohort study. Ann. Rheum. Dis., 2014;73:1914.

222. Wesley A, Bengtsson C, Elkan A-C, Klareskog L, Alfredsson L *et al.* Association between body mass index and anti–citrullinated protein antibody–positive and anti–citrullinated protein antibody–negative rheumatoid arthritis: Results from a population-based case–control study. Arthritis Care Res., 2012;65:107.
223. Maxwell JR, Gowers IR, Moore DJ, Wilson AG. Alcohol consumption is inversely associated with risk and severity of rheumatoid arthritis. Rheumatology, 2010;49:2140.
224. Källberg H, Jacobsen S, Bengtsson C, Pedersen M, Padyukov L *et al.* Alcohol consumption is associated with decreased risk of rheumatoid arthritis; Results from two Scandinavian case-control studies. Ann. Rheum. Dis., 2009;68:222.
225. Pischon N, Pischon T, Kröger J, Gülmez E, Kleber B-M *et al.* Association Among Rheumatoid Arthritis, Oral Hygiene, and Periodontitis. J. Periodontol., 2008;79:979.
226. Mercado FB, Marshall RI, Klestov AC, Bartold PM. Relationship Between Rheumatoid Arthritis and Periodontitis. J. Periodontol., 2001;72:779.
227. Scher JU, Sczesnak A, Longman RS, Segata N, Ubeda C *et al.* Expansion of intestinal Prevotella copri correlates with enhanced susceptibility to arthritis. eLife, 2013;2:e01202.
228. Scher JU, Joshua V, Artacho A, Abdollahi-Roodsaz S, Öckinger J *et al.* The lung microbiota in early rheumatoid arthritis and autoimmunity. Microbiome, 2016;4:60.
229. Tan EM, Smolen JS. Historical observations contributing insights on etiopathogenesis of rheumatoid arthritis and role of rheumatoid factor. J. Exp. Med., 2016;213:1937.
230. Nakano K, Whitaker JW, Boyle DL, Wang W, Firestein GS. DNA methylome signature in rheumatoid arthritis. Ann. Rheum. Dis., 2013;72:110.
231. Karouzakis E, Rengel Y, Jüngel A, Kolling C, Gay RE *et al.* DNA methylation regulates the expression of CXCL12 in rheumatoid arthritis synovial fibroblasts. Genes Immun., 2011;12:643.
232. Nile CJ, Read RC, Akil M, Duff GW, Wilson AG. Methylation status of a single CpG site in the IL6 promoter is related to IL6 messenger RNA levels and rheumatoid arthritis. Arthritis Rheum., 2008;58:2686.
233. Ai R, Whitaker JW, Boyle DL, Tak PP, Gerlag DM *et al.* DNA Methylome Signature in Synoviocytes From Patients With Early Rheumatoid Arthritis Compared to Synoviocytes From Patients With Longstanding Rheumatoid Arthritis. Arthritis Rheumatol., 2015;67:1978.
234. Ai R, Hammaker D, Boyle DL, Morgan R, Walsh AM *et al.* Joint-specific DNA methylation and transcriptome signatures in rheumatoid arthritis identify distinct pathogenic processes. Nat. Commun., 2016;7:11849.
235. Meng W, Zhu Z, Jiang X, Too CL, Uebe S *et al.* DNA methylation mediates genotype and smoking interaction in the development of anti-citrullinated peptide antibody-positive rheumatoid arthritis. Arthritis Res. Ther., 2017;19:71.
236. Chung Y-L, Lee M-Y, Wang A-J, Yao L-F. A therapeutic strategy uses histone deacetylase inhibitors to modulate the expression of genes involved in the pathogenesis of rheumatoid arthritis. Mol. Ther., 2003;8:707.

237. Lin HS, Hu CY, Chan HY, Liew YY, Huang HP *et al.* Anti-rheumatic activities of histone deacetylase (HDAC) inhibitors in vivo in collagen-induced arthritis in rodents. Br. J. Pharmacol., 2007;150:862.
238. Makrygiannakis D, Hermansson M, Ulfgren AK, Nicholas AP, Zendman AJW *et al.* Smoking increases peptidylarginine deiminase 2 enzyme expression in human lungs and increases citrullination in BAL cells. Ann. Rheum. Dis., 2008;67:1488.
239. McGraw WT, Potempa J, Farley D, Travis J. Purification, Characterization, and Sequence Analysis of a Potential Virulence Factor from Porphyromonas gingivalis, Peptidylarginine Deiminase. Infect. Immun., 1999;67:3248.
240. Rantapää-Dahlqvist S, de Jong BAW, Berglin E, Hallmans G, Wadell G *et al.* Antibodies against cyclic citrullinated peptide and IgA rheumatoid factor predict the development of rheumatoid arthritis. Arthritis Rheum., 2003;48:2741.
241. Buch MH, Boyle DL, Rosengren S, Saleem B, Reece RJ *et al.* Mode of action of abatacept in rheumatoid arthritis patients having failed tumour necrosis factor blockade: a histological, gene expression and dynamic magnetic resonance imaging pilot study. Ann. Rheum. Dis., 2009;68:1220.
242. Kremer JM, Russell AS, Emery P, Abud-Mendoza C, Szechinski J *et al.* Long-term safety, efficacy and inhibition of radiographic progression with abatacept treatment in patients with rheumatoid arthritis and an inadequate response to methotrexate: 3-year results from the AIM trial. Ann. Rheum. Dis., 2011;70:1826.
243. Wong PKK, Quinn JMW, Sims NA, van Nieuwenhuijze A, Campbell IK *et al.* Interleukin-6 modulates production of T lymphocyte–derived cytokines in antigen-induced arthritis and drives inflammation-induced osteoclastogenesis. Arthritis Rheum., 2005;54:158.
244. Kotake S, Udagawa N, Takahashi N, Matsuzaki K, Itoh K *et al.* IL-17 in synovial fluids from patients with rheumatoid arthritis is a potent stimulator of osteoclastogenesis. J. Clin. Invest., 1999;103:1345.
245. Buckner JH. Mechanisms of impaired regulation by CD4(+)CD25(+)FOXP3(+) regulatory T cells in human autoimmune diseases. Nat. Rev. Immunol., 2010;10:849.
246. Trynka G, Sandor C, Han B, Xu H, Stranger BE *et al.* Chromatin marks identify critical cell types for fine mapping complex trait variants. Nat. Genet., 2013;45:10.1038/ng.2504.
247. Schmidt D, Goronzy JJ, Weyand CM. CD4+ CD7- CD28- T cells are expanded in rheumatoid arthritis and are characterized by autoreactivity. J. Clin. Invest., 1996;97:2027.
248. Fasth AER, Cao D, Van Vollenhoven R, Trollmo C, Malmström V. CD28nullCD4+ T Cells – Characterization of an Effector Memory T-Cell Population in Patients with Rheumatoid Arthritis. Scand. J. Immunol., 2004;60:199.
249. Svensson J, Arkema EV, Lundberg IE, Holmqvist M. Incidence and prevalence of idiopathic inflammatory myopathies in Sweden: a nationwide population-based study. Rheumatology, 2017;56:802.

250. Barnabe C, Joseph L, Bélisle P, Labrecque J, Barr SG *et al.* Prevalence of autoimmune inflammatory myopathy in the first nations population of Alberta, Canada. Arthritis Care Res., 2012;64:1715.
251. Dobloug C, Garen T, Bitter H, Stjärne J, Stenseth G *et al.* Prevalence and clinical characteristics of adult polymyositis and dermatomyositis; data from a large and unselected Norwegian cohort. Ann. Rheum. Dis., 2015;74:1551.
252. Arahata K, Engel AG. Monoclonal antibody analysis of mononuclear cells in myopathies. I: Quantitation of subsets according to diagnosis and sites of accumulation and demonstration and counts of muscle fibers invaded by T cells. Ann. Neurol., 1984;16:193.
253. Engel AG, Arahata K. Mononuclear cells in myopathies: Quantitation of functionally distinct subsets, recognition of antigen-specific cell-mediated cytotoxicity in some diseases, and implications for the pathogenesis of the different inflammatory myopathies. Hum. Pathol., 1986;17:704.
254. Fasth AER, Dastmalchi M, Rahbar A, Salomonsson S, Pandya JM *et al.* T Cell Infiltrates in the Muscles of Patients with Dermatomyositis and Polymyositis Are Dominated by CD28null T Cells. J. Immunol., 2009;183:4792.
255. Dalakas MC. Polymyositis, Dermatomyositis, and Inclusion-Body Myositis. N. Engl. J. Med., 1991;325:1487.
256. Lundberg IE, Tjärnlund A, Bottai M, Werth VP, Pilkington C *et al.* 2017 European League Against Rheumatism/American College of Rheumatology classification criteria for adult and juvenile idiopathic inflammatory myopathies and their major subgroups. Ann. Rheum. Dis., 2017;69:2271.
257. Bohan A, Peter JB. Polymyositis and Dermatomyositis. N. Engl. J. Med., 1975;292:344.
258. Bohan A, Peter JB. Polymyositis and Dermatomyositis. N. Engl. J. Med., 1975;292:403.
259. Brouwer R, Hengstman G, Egberts W, Ehrfeld H, Bozic B *et al.* Autoantibody profiles in the sera of European patients with myositis. Ann. Rheum. Dis., 2001;60:116.
260. Ghirardello A, Borella E, Beggio M, Franceschini F, Fredi M *et al.* Myositis autoantibodies and clinical phenotypes. Auto Immun. Highlights, 2014;5:69.
261. Zampieri S, Ghirardello A, Iaccarino L, Tarricone E, Gambari PF *et al.* Anti-Jo-1 Antibodies. Autoimmunity, 2005;38:73.
262. Stone KB, Oddis CV, Fertig N, Katsumata Y, Lucas M *et al.* Anti–Jo-1 antibody levels correlate with disease activity in idiopathic inflammatory myopathy. Arthritis Rheum., 2007;56:3125.
263. Greenberg SA, Pinkus JL, Pinkus GS, Burleson T, Sanoudou D *et al.* Interferon-α/β–mediated innate immune mechanisms in dermatomyositis. Ann. Neurol., 2005;57:664.
264. Walsh RJ, Kong SW, Yao Y, Jallal B, Kiener PA *et al.* Type I Interferon–Inducible Gene Expression in Blood Is Present and Reflects Disease Activity in Dermatomyositis and Polymyositis. Arthritis Rheum., 2007;56:3784.
265. Hirsch TJ, Enlow RW, Bias WB, Arnett FC. HLA-D related (DR) antigens in various kinds of myositis. Hum. Immunol., 1981;3:181.

266. O'Hanlon TP, Carrick DM, Arnett FC, Reveille JD, Carrington M *et al.* Immunogenetic Risk and Protective Factors for the Idiopathic Inflammatory Myopathies: Distinct HLA-A, -B, -Cw, -DRB1 and -DQA1 Allelic Profiles and Motifs Define Clinicopathologic Groups in Caucasians. Medicine, 2005;84:338.
267. Miller FW, Chen W, O'Hanlon TP, Cooper RG, Vencovsky J *et al.* Genome-wide Association Study Identifies HLA 8.1 Ancestral Haplotype Alleles as Major Genetic Risk Factors for Myositis Phenotypes. Genes Immun., 2015;16:470.
268. Hengstman G, van Venrooij WJ, Vencovsky J, Moutsopoulos H, van Engelen BGM. The relative prevalence of dermatomyositis and polymyositis in Europe exhibits a latitudinal gradient. Ann. Rheum. Dis., 2000;59:141.
269. Azali P, Barbasso Helmers S, Kockum I, Olsson T, Alfredsson L *et al.* Low serum levels of vitamin D in idiopathic inflammatory myopathies. Ann. Rheum. Dis., 2013;72:512.
270. Chinoy H, Adimulam S, Marriage F, New P, Vincze M *et al.* Interaction of HLA-DRB1*03 and smoking for the development of anti-Jo-1 antibodies in adult idiopathic inflammatory myopathies: a European-wide case study. Ann. Rheum. Dis., 2012;71:961.
271. Hill CL, Zhang Y, Sigurgeirsson B, Pukkala E, Mellemkjaer L *et al.* Frequency of specific cancer types in dermatomyositis and polymyositis: a population-based study. Lancet, 2001;357:96.
272. Goebels N, Michaelis D, Engelhardt M, Huber S, Bender A *et al.* Differential expression of perforin in muscle-infiltrating T cells in polymyositis and dermatomyositis. J. Clin. Invest., 1996;97:2905.
273. Pandya JM, Venalis P, Al-Khalili L, Shahadat Hossain M, Stache V *et al.* CD4+ and CD8+ CD28null T Cells Are Cytotoxic to Autologous Muscle Cells in Patients With Polymyositis. Arthritis Rheumatol., 2016;68:2016.
274. Arnett FC, Edworthy SM, Bloch DA, McShane DJ, Fries JF *et al.* The American Rheumatism Association 1987 revised criteria for the classification of rheumatoid arthritis. Arthritis Rheum., 1988;31:315.
275. Liu Y, Aryee MJ, Padyukov L, Fallin MD, Hesselberg E *et al.* Epigenome-wide association data implicate DNA methylation as an intermediary of genetic risk in Rheumatoid Arthritis. Nat. Biotechnol., 2013;31:142.
276. Folkersen L, Brynedal B, Diaz-Gallo LM, Ramsköld D, Shchetynsky K *et al.* Integration of Known DNA, RNA and Protein Biomarkers Provides Prediction of Anti-TNF Response in Rheumatoid Arthritis: Results from the COMBINE Study. Mol. Med., 2016;22:322.
277. Kim D, Pertea G, Trapnell C, Pimentel H, Kelley R *et al.* TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. Genome Biol., 2013;14:R36.
278. Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics, 2010;26:139.
279. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C *et al.* STAR: ultrafast universal RNA-seq aligner. Bioinformatics, 2013;29:15.

280. Ramsköld D, Wang ET, Burge CB, Sandberg R. An Abundance of Ubiquitously Expressed Genes Revealed by Tissue Transcriptome Sequence Data. PLoS Comput. Biol., 2009;5:e1000598.
281. Horton R, Gibson R, Coggill P, Miretti M, Allcock RJ *et al.* Variation analysis and gene annotation of eight MHC haplotypes: The MHC Haplotype Project. Immunogenetics, 2008;60:1.
282. Lee W, Plant K, Humburg P, Knight JC. AltHapAlignR: improved accuracy of RNA-seq analyses through the use of alternative haplotypes. Bioinformatics, 2018;34:2401.
283. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol., 2014;15:550.
284. Schmittgen TD, Livak KJ. Analyzing real-time PCR data by the comparative CT method. Nat. Protoc., 2008;3:1101.
285. Olerup O, Zetterquist H. HLA-DR typing by PCR amplification with sequence-specific primers (PCR-SSP) in 2 hours: an alternative to serological DR typing in clinical practice including donor-recipient matching in cadaveric transplantation. Tissue Antigens, 1992;39:225.
286. Boegel S, Löwer M, Schäfer M, Bukur T, de Graaf J *et al.* HLA typing from RNA-Seq sequence reads. Genome Med., 2012;4:102.
287. Burton PR, Clayton DG, Cardon LR, Craddock N, Deloukas P *et al.* Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature, 2007;447:661.
288. Barrett JC, Clayton DG, Concannon P, Akolkar B, Cooper JD *et al.* Genome-wide association study and meta-analysis find that over 40 loci affect risk of type 1 diabetes. Nat. Genet., 2009;41:703.
289. Okada Y, Terao C, Ikari K, Kochi Y, Ohmura K *et al.* Meta-analysis identifies nine new loci associated with rheumatoid arthritis in the Japanese population. Nat. Genet., 2012;44:511.
290. Ward LD, Kellis M. HaploReg: a resource for exploring chromatin states, conservation, and regulatory motif alterations within sets of genetically linked variants. Nucleic Acids Res., 2012;40:D930.
291. Boyle AP, Hong EL, Hariharan M, Cheng Y, Schaub MA *et al.* Annotation of functional variation in personal genomes using RegulomeDB. Genome Res., 2012;22:1790.
292. Schofield EC, Carver T, Achuthan P, Freire-Pritchett P, Spivakov M *et al.* CHiCP: a web-based tool for the integrative and interactive visualization of promoter capture Hi-C datasets. Bioinformatics, 2016;32:2511.
293. Wang J, Lee S, Teh CE-Y, Bunting K, Ma L *et al.* The transcription repressor, ZEB1, cooperates with CtBP2 and HDAC1 to suppress IL-2 gene activation in T cells. Int. Immunol., 2009;21:227.
294. Katsuyama E, Yan M, Watanabe KS, Matsushima S, Yamamura Y *et al.* Downregulation of miR-200a-3p, Targeting CtBP2, Is Involved in the Hypoproduction of IL-2 in Systemic Lupus Erythematosus–Derived T Cells. J. Immunol., 2017;198:4268.
295. Mi H, Huang X, Muruganujan A, Tang H, Mills C *et al.* PANTHER version 11: expanded annotation data from Gene Ontology and Reactome pathways, and data analysis tool enhancements. Nucleic Acids Res., 2017;45:D183.

296. Asmal M, Colgan J, Naef F, Yu B, Lee Y *et al.* Production of Ribosome Components in Effector CD4+ T Cells Is Accelerated by TCR Stimulation and Coordinated by ERK-MAPK. Immunity, 2003;19:535.
297. Micheau O, Thome M, Schneider P, Holler N, Tschopp J *et al.* The Long Form of FLIP Is an Activator of Caspase-8 at the Fas Death-inducing Signaling Complex. J. Biol. Chem., 2002;277:45162.
298. Zhang N, Hopkins K, He Y-W. c-FLIP protects mature T lymphocytes from TCR-mediated killing. J. Immunol., 2008;181:5368.
299. Vallejo AN, Schirmer M, Weyand CM, Goronzy JJ. Clonality and Longevity of $CD4^{+}CD28^{null}$ T Cells Are Associated with Defects in Apoptotic Pathways. J. Immunol., 2000;165:6301.
300. Maly K, Schirmer M. The story of $CD4^{+}$ $CD28^{-}$ T cells revisited: solved or still ongoing? J. Immunol. Res., 2015;2015:348746.
301. Ward-Kavanagh Lindsay K, Lin Wai W, Šedý John R, Ware Carl F. The TNF Receptor Superfamily in Co-stimulating and Co-inhibitory Responses. Immunity, 2016;44:1005.
302. Qui HZ, Hagymasi AT, Bandyopadhyay S, St. Rose M-C, Ramanarasimhaiah R *et al.* CD134 Plus CD137 Dual Costimulation Induces Eomesodermin in CD4 T Cells To Program Cytotoxic Th1 Differentiation. J. Immunol., 2011;187:3555.
303. Curran MA, Geiger TL, Montalvo W, Kim M, Reiner SL *et al.* Systemic 4-1BB activation induces a novel T cell phenotype driven by high expression of Eomesodermin. J. Exp. Med., 2013;210:743.
304. Knox JJ, Cosma GL, Betts MR, McLane LM. Characterization of T-Bet and Eomes in Peripheral Human Immune Cells. Front. Immunol., 2014;5:217.
305. Eshima K, Chiba S, Suzuki H, Kokubo K, Kobayashi H *et al.* Ectopic expression of a T-box transcription factor, eomesodermin, renders CD4+ Th cells cytotoxic by activating both perforin- and FasL-pathways. Immunol. Lett., 2012;144:7.
306. Pearce EL, Mullen AC, Martins GA, Krawczyk CM, Hutchins AS *et al.* Control of Effector $CD8^{+}$ T Cell Function by the Transcription Factor Eomesodermin. Science, 2003;302:1041.
307. Jung HW, Choi SW, Choi JIL, Kwon BS. Serum concentrations of soluble 4-1BB and 4-1BB ligand correlated with the disease severity in rheumatoid arthritis. Exp. Mol. Med., 2004;36:13.
308. McInnes IB, Liew FY. Interleukin 15: a proinflammatory role in rheumatoid arthritis synovitis. Immunol. Today, 1998;19:75.
309. Thurkow EW, van der Heijden IM, Breedveld FC, Smeets TJM, Daha MR *et al.* Increased expression of IL-15 in the synovium of patients with rheumatoid arthritis compared with patients with Yersina-induced arthritis and osteoarthritis. J. Pathol., 1997;181:444.
310. Fogdell A, Hillert J, Sachs C, Olerup O. The multiple sclerosis- and narcolepsy-associated HLA class II haplotype includes the DRB5*0101 allele. Tissue Antigens, 1995;46:333.
311. Kerlan-Candon S, Combe B, Vincent R, Clot J, Pinet V *et al.* HLA-DRB1 gene transcripts in rheumatoid arthritis. Clin. Exp. Immunol., 2001;124:142.

312. Rao DA, Gurish MF, Marshall JL, Slowikowski K, Fonseka CY *et al.* Pathologically expanded peripheral T helper cell subset drives B cells in rheumatoid arthritis. Nature, 2017;542:110.
313. Fonseka CY, Rao DA, Teslovich NC, Korsunsky I, Hannes SK *et al.* Mixed-effects association of single cells identifies an expanded effector $CD4^+$ T cell subset in rheumatoid arthritis. Sci. Transl. Med., 2018;10:eaaq0305.
314. Baecher-Allan C, Wolf E, Hafler DA. MHC Class II Expression Identifies Functionally Distinct Human Regulatory T Cells. J. Immunol., 2006;176:4622.
315. Dalakas MC, Hohlfeld R. Polymyositis and dermatomyositis. Lancet, 2003;362:971.
316. Engel AG, Arahata K. Monoclonal antibody analysis of mononuclear cells in myopathies. II: Phenotypes of autoinvasive cells in polymyositis and inclusion body myositis. Ann. Neurol., 1984;16:209.
317. Orimo S, Koga R, Goto K, Nakamura K, Arai M *et al.* Immunohistochemical analysis of perforin and granzyme A in inflammatory myopathies. Neuromuscul. Disord., 1994;4:219.
318. Aran D, Hu Z, Butte AJ. xCell: digitally portraying the tissue cellular heterogeneity landscape. Genome Biol., 2017;18:220.
319. Li J, Mahajan A, Tsai M-D. Ankyrin Repeat: A Unique Motif Mediating Protein−Protein Interactions. Biochemistry, 2006;45:15168.
320. Stahl EA, Raychaudhuri S, Remmers EF, Xie G, Eyre S *et al.* Genome-wide association study meta-analysis identifies seven new rheumatoid arthritis risk loci. Nat. Genet., 2010;42:508.
321. Jostins L, Ripke S, Weersma RK, Duerr RH, McGovern DP *et al.* Host-microbe interactions have shaped the genetic architecture of inflammatory bowel disease. Nature, 2012;491:119.
322. Alloza I, Otaegui D, de Lapuente AL, Antigüedad A, Varadé J *et al.* ANKRD55 and DHCR7 are novel multiple sclerosis risk loci. Genes Immun., 2011;13:253.
323. Zimmer DB, Cornwall EH, Landar A, Song W. The S100 protein family: History, function, and expression. Brain Res. Bull., 1995;37:417.
324. Steiner J, Marquardt N, Pauls I, Schiltz K, Rahmoune H *et al.* Human CD8+ T cells and NK cells express and secrete S100B upon stimulation. Brain Behav. Immun., 2011;25:1233.
325. Hsing Lianne C, Rudensky Alexander Y. The lysosomal cysteine proteases in MHC class II antigen presentation. Immunol Rev., 2005;207:229.
326. Mackay LK, Minnich M, Kragten NAM, Liao Y, Nota B *et al.* Hobit and Blimp1 instruct a universal transcriptional program of tissue residency in lymphocytes. Science, 2016;352:459.
327. van Gisbergen KPJM, Kragten NAM, Hertoghs KML, Wensveen FM, Jonjic S *et al.* Mouse Hobit is a homolog of the transcriptional repressor Blimp-1 that regulates NKT cell effector differentiation. Nat. Immunol., 2012;13:864.
328. Oja AE, Vieira Braga FA, Remmerswaal EBM, Kragten NAM, Hertoghs KML *et al.* The Transcription Factor Hobit Identifies Human Cytotoxic CD4(+) T Cells. Front. Immunol., 2017;8:325.

329. Albrecht I, Wick C, Hallgren Å, Tjärnlund A, Nagaraju K *et al.* Development of autoantibodies against muscle-specific FHL1 in severe inflammatory myopathies. J. Clin. Invest., 2015;125:4612.
330. Levine Stuart M, Raben N, Xie D, Askin Frederic B, Tuder R *et al.* Novel conformation of histidyl–transfer RNA synthetase in the lung. Arthritis Rheum., 2007;56:2729.
331. Confalonieri P, Bernasconi P, Cornelio F, Mantegazza R. Transforming Growth Factor-β1 in Polymyositis and Dermatomyositis Correlates with Fibrosis but not with Mononuclear Cell Infiltrate. J. Neuropathol. Exp. Neurol., 1997;56:479.
332. Dominguez CX, Amezquita RA, Guan T, Marshall HD, Joshi NS *et al.* The transcription factors ZEB2 and T-bet cooperate to program cytotoxic T cell terminal differentiation in response to LCMV viral infection. J. Exp. Med., 2015;212:2041.
333. Omilusik KD, Best JA, Yu B, Goossens S, Weidemann A *et al.* Transcriptional repressor ZEB2 promotes terminal differentiation of CD8+ effector and memory T cell populations during infection. J. Exp. Med., 2015;212:2027.
334. Verschueren K, Remacle JE, Collart C, Kraft H, Baker BS *et al.* SIP1, a Novel Zinc Finger/Homeodomain Repressor, Interacts with Smad Proteins and Binds to 5′-CACCT Sequences in Candidate Target Genes. J. Biol. Chem., 1999;274:20489.
335. Englund P, Wahlström J, Fathi M, Rasmussen E, Grunewald J *et al.* Restricted T cell receptor BV gene usage in the lungs and muscles of patients with idiopathic inflammatory myopathies. Arthritis Rheum., 2007;56:372.