Assessment of the role of the thymic transcriptome in genetic predisposition to autoimmune diseases

Thesis for the degree of Philosophiae Doctor (PhD)

By

Ingvild Synnøve Matre Gabrielsen

2018

Department of Medical Genetics, Oslo University Hospital, Ullevål, Oslo

Faculty of Medicine, University of Oslo

Series of dissertations submitted to the
Faculty of Medicine, University of Oslo

ISBN 978-82-8377-399-6

Cover: Hanne Baadsgaard Utigard.
Print production: Reprosentralen, University of Oslo.

Table of Contents

Acknowledgement

This work has been carried out at the Department of Medical Genetics, Oslo University Hospital Ullevål and University of Oslo during the period 2012-2018. I would like to thank the head of the Department, Dag Undlien, for an excellent research environment.

I am very thankful for the research grant I received from the Norwegian research council.

I am especially grateful to my eminent, main supervisor **Benedicte A. Lie** for her guidance through the complex field of genetics. Thank you Benedicte, for sharing so much of your impressive knowledge and giving me so many good advices (the best one was definitely "one thing at a time, as Alfie Atkins father said"). You are always available, optimistic and have so much enthusiasm, which really encouraged and motivated me to carry on. I feel privileged to have been part of your research group these five years.

Siri T. Flåm, you deserve a special thanks, you are the Duracell bunny of the group. You have an impressing working capacity and have helped me so much with the thymic tissue in the lab which has required many late nights. **Hanna Helgeland**, you deserve special thanks too, for all the programming you did in these projects. I learned so much from studying your scripts; it really opened my eyes for programming.

My amazing colleagues, **Marte K. Viken**, **Silja S. Amundsen**, **Arvind Sundaram, Hans Christian Dalsbotten Aass**, **Helle Akselsen**, and **Kristian Holm** thank you for sharing your knowledge, your participation in scientific discussions, all your help in the lab, with the data and all the advice you have given me in the projects and in the manuscripts.

I am also very grateful to Harald Lindberg, all the patients, Tone and all the nurses and surgeons at Rikshospitalet who provided us with the thymic tissues.

A big thanks to all the other members of my group who always contribute to such a nice working environment: **Kari Guderud**, **Line Sunde**, **Fatima Heinicke**, **Asgeir Lande**, **Eva Kristine Henriksen**, **Maria Vigeland** and **Fatemeh Kaveh**.

Lastly, thanks to **Ylva Kaiser** and **Cara Bray** who made me understand that scientific conference is not only about learning, but also about networking and making friends.

Finally, I would like to thank my **family** and **friends**, and especially to **Svenn-Arne Dragly.** Thank you for all those days you got up early in the morning and drove me, first to Ullevål to get chemicals, and then to Rikshospitalet to get thymic tissues before 8:00 am. Thank you for driving up to Jessheim during my last weeks of writing before the deadline to buy the last Air Conditioner they had, when the whole Oslo was sold out and the temperature in my office was passing 35°C. You are always so loving, helpful and supporting.

Ingvild Gabrielsen

Oslo, 2018

Abbreviations

AH	autoimmune hepatitis
AID	autoimmune disease
AIRE	autoimmune regulator
APC	antigen presenting cell
ATD	autoimmune thyroid disease
BCR	B cell receptor
CD	Crohn’s disease
CD123	Cluster of differentiation 123
CD141	Cluster of differentiation 141
CeD	Celiac disease
CPM	count-per-million
cTEC	cortical thymic epithelial cell
DC	dendritic cell
DN	double negative
DNA	deoxyribonucleic acid
DP	double positive
ENCODE	Encyclopedia of DNA elements
FDR	False discovery rate
FPKM	Fragments per Kilobase of transcript per Million mapped reads
eProbe	eQTL Probe
eQTL	expression quantitative trait locus
eSNP	eQTL SNP
GD	Grave´s disease
GWAS	genome-wide association studies
HLA	human leukocyte antigen
IBD	inflammatory bowel disease
IQR	Interquartile range
JIA	juvenile idiopathic arthritis
LD	linkage disequilibrium
lincRNA	long intergenic non-coding RNA

lncRNA	long non-coding RNA
MAF	minor allele frequency
MDS	multidimensional scaling plot
MG	myasthenia gravis
mRNA	messenger RNA
MS	multiple sclerosis
mTEC	medullary thymic epithelial cell
NHGRI	National Human Genome Research Institute
NIH	National Institutes of Health
NSC	Norwegian Sequencing Center
PBC	primary biliary cirrhosis
Ps	Psoriasis
PSC	primary sclerosing cholangitis
RA	rheumatoid arthritis
RNA	ribonucleic acid
SLE	systemic lupus erythematosus
SP	single positive
SNP	single nucleotide polymorphism
Std. RNA-Seq	standard RNA sequencing
T1D	type 1 diabetes
TCR	T cell receptors
TPM	transcript per million
TRA	tissue-restricted antigen
Treg	regulatory T cell
UC	ulcerative colitis
UTR	untranslated region

List of publications

Paper I

Gabrielsen ISM, Viken MK, Amundsen SS, Helgeland H, Holm K, Flåm ST, Lie BA. Autoimmune risk variants in ERAP2 are associated with gene-expression levels in thymus. Genes and immunity 2016. Volume 17, pages: 406-411.

Paper II

Gabrielsen IS, Amundsen SS, Helgeland H, Flåm ST, Hatinoor N, Holm K, Viken MK, Lie BA. Genetic risk variants for autoimmune diseases that influence gene expression in thymus. Human molecular genetics 2016. Volume 25, pages: 3117-3124.

Paper III

Gabrielsen IS, Helgeland H, Akselsen H, Aass HC, Sundaram AYM., Snowhite I, Pugliese A, Flåm ST, Lie BA. Transcriptomes of antigen presenting cells in human thymus.

(Submitted)

1. Introduction

1.1 Autoimmune diseases: general features and epidemiology

Autoimmune diseases (AIDs) comprise a diversity of progressive, chronic inflammatory disorders where the immune system attacks the body’s own tissues and organs (Figure 1). Today, the prevalence of AIDs is estimated to be approximately 7 - 9 % [1] in the developed world, with type 1 diabetes (T1D) and autoimmune thyroid disease (ATD) being the most common disorders [2]. Nearly 100 different AIDs are known, and several diseases share common symptoms, like malaise, fatigue, fever, joint pain and rash. Co-occurrence of several AIDs are also often observed within one individual, making it difficult to diagnose some of these patients correctly [3]. Currently, autoimmunity is treated by relieving symptoms, as no curative therapy yet exists.

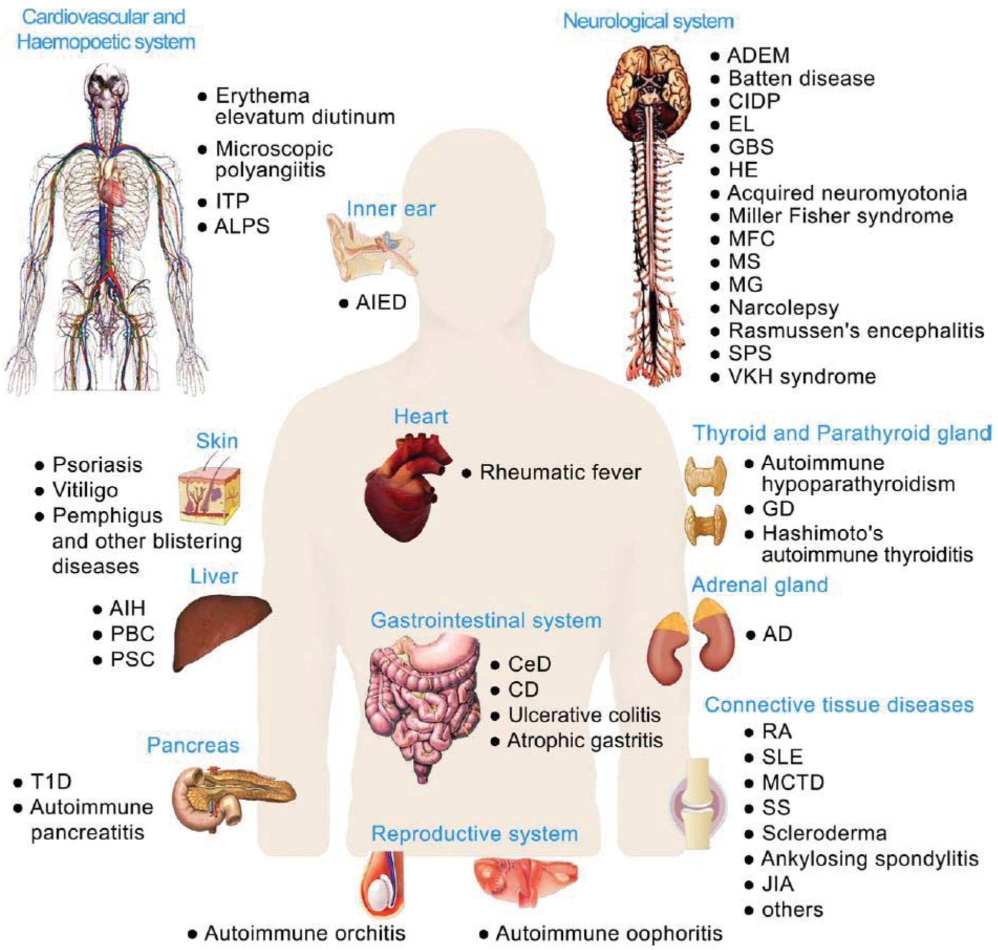

Figure 1: The vast range of AIDs affects a large number of organs and tissues (Figure from [2] with permission).

Although many AIDs exist today, more epidemiological data is available for the most well-known diseases. Geographically, we know that the prevalence rates for 29 AIDs (including seven listed in Table 1) span overlapping ranges across countries [1]. These include type 1 diabetes (T1D), primary biliary cirrhosis (PBC), Crohn's disease (CD) and systemic lupus erythematosus (SLE). However, for multiple sclerosis (MS), autoimmune hepatitis (AH) and ulcerative colitis (UC), the prevalence rates are reported to be higher in Europe, North America, Australia and New Zealand compared to Asia, Middle East, Caribbean and South America. Incidence rates for AIDs report that MS, T1D, PBC, AH, CD, UC and SLE are higher in North America and Europe compared to Asia and the Middle East [4].

AIDs also differ according to ethnicity, age at onset and gender. For example, in the United States, where different ethnic groups live in the same area, blacks are at higher risk for developing SLE and systemic sclerosis (SSc) compared with whites [5]. In contrast, T1D has a lower incidence rate among blacks and Hispanics. Also, blacks and Asians have a lower risk of developing MS. White, Hispanics and blacks have similar rates for rheumatoid arthritis (RA). The mean age of onset in childhood T1D and juvenile idiopathic arthritis (JIA) is 8-10 years [5] and around the age of 12 in juvenile myasthenia gravis (MG) [6]. Adult MG, MS and Grave´s disease (GD) generally occur between the ages 30-50, whereas AIDs with later age of onset (40–70 years) include myositis, thyroiditis, Sjögren disease and RA. Most AIDs affect women more frequently, and are among the leading cause of death in the United States for young and middle aged women under 65 years of age [7]. The exceptions are CD (1.2:1 male to female ratio) and primary sclerosing cholangitis (PSC) (2:1), where the prevalence is higher in men. The overrepresentation of various autoimmune disorders in women could indicate that a hormonal influence is implicated in the development of autoimmunity. Skewed X-inactivation has also been proposed as a mechanism associated with AIDs in women (reviewed in [8]).

Table 1. Prevalence data for seven AIDs, adapted from [1] with permission. This table includes both hospitalized and non-hospitalized data.

	Studies from Europe, North America, Australia, New Zealand		Studies from Asia, Middle East, Caribbean, South America	
Disease	Rate per 100,000	Study Area	Rate per 100,000	Study area
Multiple sclerosis	182 177-358 121-200 46 50	Denmark US, Canada Italy, Greece, France, Ireland Norway Portugal, New Zealand	4-20 13 11-62 101	Colombia, Brazil, Argentina Japan Israel, Kuwait, Jordan, Iran Turkey
Type 1 Diabetes (all ages)	946 118 340-570	Denmark Lithuania UK, Sweden, Australia	-	-
Type 1 Diabetes (< 20 years of age)	87-120 227-335 70	Spain, Germany US, New Zealand US – American Indian	31 110-270	Bahamas Kuwait, Saudi Arabia
Primary biliary cirrhosis	12 15-40 4-20	Denmark Norway, Finland, Spain, UK US, Australia	4-18	Israel
Autoimmune hepatitis	45 11-17 36	Denmark Spain, Sweden, Norway US – Alaska Natives	3-8	Singapore
Crohn´s disease	225 28-53	Denmark Bosnia-Herzegovina, Hungary	6-53	Puerto Rico, Malaysia, Lebanon
Ulcerative colitis	378 143-294	Denmark US, Hungary, Denmark, New Zealand	6 102	Lebanon Puerto Rico
Systemic lupus erythematosus	32 10-66 34-150 42	Denmark US – Native Americans US, Spain, Greece Canada – 1st Nations	30 19 45 93	Philipines Saudi Arabia Australia Australia – aboriginal

1.2 Pathogenesis

The immune system defends the organism against foreign invaders, and can be divided into two branches: innate and adaptive immunity. The innate immune system is non-specific and is mainly composed of the skin, epithelial and mucosal linings of the gastrointestinal tract and different types of immune cells, such as phagocytes, dendritic cells (DCs) and natural killer cells. The adaptive immunity is an acquired defense mechanism that consists of T and B cells. B cells can bind to antigens, i.e. proteins capable of inducing a specific immune response, with their B cell receptor (BCR). Activation of B cells leads to secretion of antibodies specifically directed against this antigen. Similarly, T cells have T cell receptors (TCR) that bind to antigens. However, for TCR recognition, the antigen must be presented in conjunction with a human leukocyte antigen (HLA)-molecule. T cells can distinguish between foreign peptides originating from e.g. bacteria or viruses and self-peptides, i.e. peptides from the cells and organs in the body. The non-responsiveness to self-peptides is due to the body's tolerance mechanisms, and enables the efficient removal of pathogens without harming the organism.

Although the body has both central and peripheral tolerance mechanisms, autoimmune reactions (i.e reactions to self) do occur. In fact, they play a part of the physiological functioning of the immune system [9]. Normally, healthy individuals have a small fraction of autoreactive T cells [10] and B cells [11] in their bloodstream. Low concentration of natural autoantibodies (i.e. self-reactive antibodies) can also be found [12]. These are usually IgM isotype antibodies, and the B cells producing these antibodies have not undergone somatic hypermutation, a characteristic of T cell-dependent adaptive immune response [9]. These natural antibodies are believed to facilitate the removal of senescent cells and autoantigens.

Autoreactive immune cells are usually strictly controlled and can be rapidly removed by immunoregulatory mechanisms [9]. However, if the sophisticated mechanisms of the immune system that regulate the maintenance of tolerance are disturbed, they might lead to the development of an AID. Disturbances proposed to be involved in the loss of tolerance can be failure to delete autoreactive lymphocytes, molecular mimicry, abnormal presentation of self-peptides, epitope spreading or polyclonal lymphocyte activation.

AIDs arise when tolerance is lost and a sustained immune response persists against one or several self-peptides. Peptides that are no longer recognized by the immune system as "self", are referred to as "autoantigens". An immunological response against autoantigens can be T- or B cell mediated, or both [9]. The autoimmune attack is usually damaging to the targeted organ, and either result in a complete loss-of-function (e.g. in T1D or Hashimoto thyroiditis), or hyperstimulation or inhibition of its function (e.g. in MS).

AIDs can be broadly divided into two major groups; systemic or organ-specific. Immune responses that are directed against specific organs or tissues are referred to as organ-specific AIDs. A well-known organ specific AID is T1D. Established autoantigens in T1D are for example insulin [13], non-specific islet cell antigens [14], insulinoma antigen-2 [15] and glutamic acid decarboxylase 65 [16]. The autoimmune response can also impair or damage several tissues at the same time (e.g. in SLE or SSc). These are referred to as systemic AIDs. These are often characterized by immune responses against a large variety of autoantigens, including deoxyribonucleic acid (DNA), cell surface molecules and intracellular proteins. In SLE, double-stranded DNA and Sm antigens of the U-1 small nuclear ribonucleoprotein complex are considered pathognomonic [17].

1.3 Etiology

The etiology of AIDs is largely unknown. Our main understanding is that AIDs develop from a combination of genetic susceptibility and environmental factors. Environmental factors are biotic or abiotic elements that can influence living organisms, such as diet, pathogens, chemicals and climate. Environmental interactions in autoimmunity have for example been seen in RA, where smoking has been identified as an important risk factor [18], or in coeliac disease, where gluten leads to pathological changes in the small intestine [19]. Other environmental risk factors with less clear roles include nutrition, the microbiota, infectious processes and xenobiotics (tobacco smoke, pharmaceutical agents, hormones, ultraviolet light, silica solvents, heavy metals, vaccines and collagen/silicone implants [2]). At the epigenetic level, mechanisms such as DNA methylation and histone modifications have been found to influence gene expression in different AIDs. Hypermethylation of CpGs at the *INS* locus has for example been found in T1D [20] whereas extensive demethylation of the *PAD2* promoter region was found in MS [21]. Histone H3 and H4 hypoacetylation has also been observed in CD4+ T cells in SLE patients [22].

Epidemiological studies of most AIDs have shown that strong heritability exists [23]. The concordance rate of AIDs in monozygotic twins ranges from 12% to 68% [2], and siblings of proband cases have an increased risk compared to the general population [23]. This indicates that there are genetic risk factors contributing to the development of autoimmunity. I will now give an introduction to the field of human genetics, and further describe how genetic, interindividual differences can affect health and disease.

1.4 Genetics of complex diseases

1.4.1 Human genetic variation and linkage disequilibrium

The human genome consists of 3.1 billion base pairs (bp) distributed across 23 chromosome pairs in the cell nucleus. The genome is estimated to harbor 19000 protein coding genes [24]. The vast majority of the genome (99.9%) is identical between any two unrelated individuals. However, 0.1% of the human genome varies between individuals, and this variation can be divided into simple nucleotide variations (SNVs) and structural variations (SVs). SNVs are smaller sequence variations including single nucleotide polymorphisms (SNPs), where only one bp differs between two homologous chromosomes, and small deletions and insertions. Larger sequence variations, or structural variants, comprise large indels (100 bp – 1 kilobasepair (kb)), copy number variations (duplications and deletions > 1 kb in size), and rearrangements such as inversions and translocations [25].

Genetic variation underlies the differences observed in individual phenotypic traits, such as eye color, height or blood type. Genetic variants can also confer susceptibility to a disease. A key goal of human genetics is to determine the genetic variants that affect phenotypic variation, including disease susceptibility, different response to drugs, treatment and public health. SNPs have been widely surveyed in many individuals for this purpose. A SNP can vary between two (or sometimes three) alleles, i.e, alternative forms of single nucleotides (Figure 2), and is classified as common if the frequency of the minor allele (MAF) is above 1% in the population. SNP allele frequencies have been mapped in many populations around the world, due to large efforts such as the HapMap project [26] and the 1000 Genomes project [27]. The combination of the two SNP alleles that an individual carries on a pair of homologous chromosomes is a genotype, and can either be homozygous (A/A or B/B) or heterozygous (A/B).

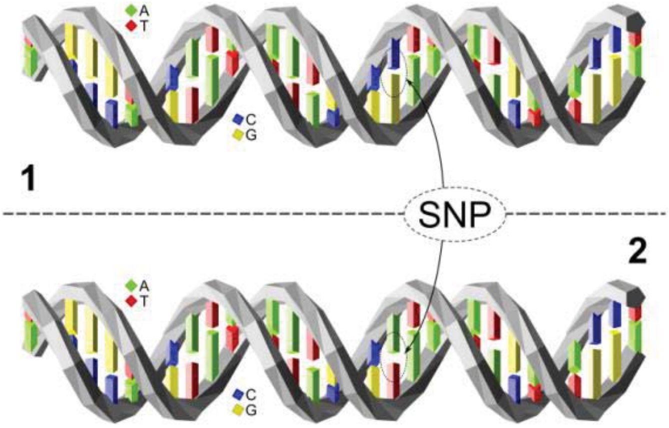

Figure 2: A SNP is a variation in a single base pair at a specific position in the genome, which can differ between alleles (SNP model by David Eccles (gringer), CC BY 4.0 https://commons.wikimedia.org/w/index.php?curid=2355125)

Furthermore, a haplotype is a sequence of multiple alleles along a single chromosome. Two SNP alleles are linked if they are transmitted together from parent to offspring more often than expected under independent inheritance and in linkage disequilibrium (LD) if, across a population as a whole, they are found on the same haplotype more often than expected. On the contrary, when two SNP alleles are independently inherited, they are in linkage equilibrium. LD can be measured with r^2, which is the square of the statistical correlation coefficient between two SNPs. This measure unit has absolute values between 0 and 1 (Figure 3), and an r^2 value of 1 means that the alleles are in perfect LD and are always inherited together in the population. Because r^2 is bidirectional, this means that the two SNPs can only give two possible haplotypes. LD is important when finding genetic association with a trait, as the causal SNP can be in LD with the nucleotide position where the association signal is detected.

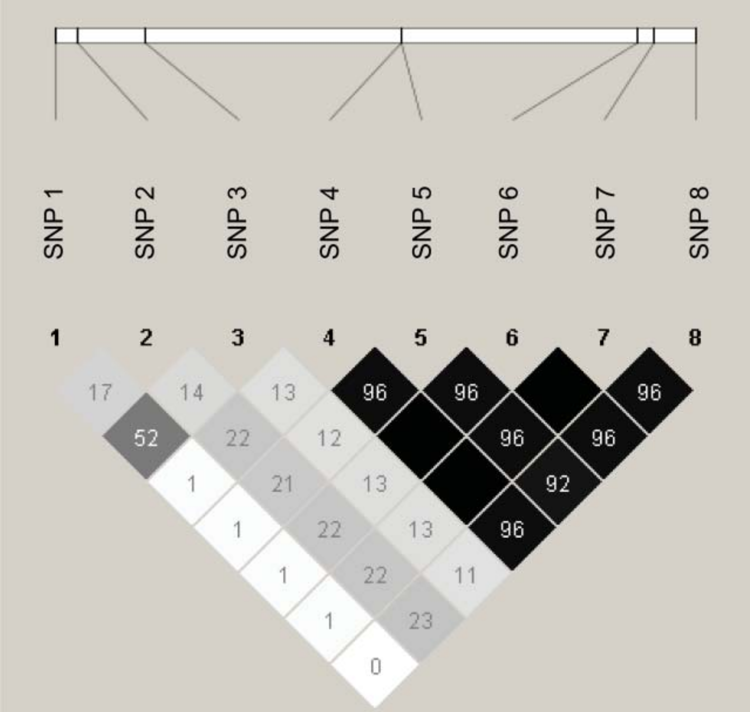

Figure 3: The LD between eight SNPs measured in r^2 show how well these SNP are correlated.

1.4.2 Genetic association studies

A tremendous advance in identifying genetic risk variants has been through genome-wide association studies (GWAS). GWAS is a hypothesis free method that enables the simultaneous genotyping of hundreds of thousands of common SNPs. GWAS compares the frequencies of SNP alleles in patients and healthy controls to examine whether any variant is associated with a trait. Significance threshold for GWAS is usually set at $P < 5 \times 10^{-8}$. However, the question of what strength of evidence should be considered significant is somewhat controversial [28]. Today, GWAS have identified thousands of genetic risk variants associated with a large variety of diseases, where hundreds are associated with immune mediated diseases (including AIDs) [29]. All GWASs can be found in the GWAS catalog (https://www.ebi.ac.uk/gwas/), which contains (as of the 12^{th} of May 2018) 3379 publications and 61620 unique SNP-trait associations.

Before GWAS, some of the first and most significant AID susceptibility loci discovered were the HLA genes [23]. The HLA genes were further confirmed in most GWAS studies focusing on AIDs. In addition, GWAS detected a large number of non-HLA loci with smaller relative risks, usually < 1.2 [23]. These are not strong genetic determinants; however, they could give an indication of biological pathways leading to disease rather than identifying etiological factors.

Between 2012 and 2015, AID association signals from GWAS were further replicated and fine-mapped in a wave of Immunochip studies. The Immunochip is a SNP genotyping array with dense marker coverage across 186 genetic regions harboring risk variants for 12 well-defined AIDs [30]. To date, the use of Immunochip has replicated previously known risk SNPs, as well as identified novel risk loci in Ankylosing Spondylitis (AS) [31], RA [32], ATD [33], Psoriasis (Ps) [34], Celiac Disease (CeD) [35], Inflammatory Bowel disease (IBD) [36], MS [37], Atopic Dermatitis (AD) [38], PSC [39], SSc [40], JIA [41] and T1D [42]. The Immunochip has narrowed the association signals in various AIDs [29]. For example in CeD, almost half of the known association signals have been narrowed to individual genes or subregions of genes [35].

GWAS and Immunochip studies led us to understand that several risk loci overlap between AIDs. Although sharing is common, it is also complex [29]. If two diseases share a SNP or a haplotype that increases the risk for both AIDs, the overlap is “correlated and concordant”. When a shared locus comprises a haplotype that increases risk for one disease but is protective for the other, it is “correlated but discordant”. Finally, if two different haplotypes are implicated, it is “non-correlated”. As an example, the six AIDs AS, CeD, IBD, Ps, RA and T1D share 71 loci at $P < 5 \times 10^{-8}$ between two or more diseases. Pairwise, 416 risk loci overlap, where 45% are correlated and concordant, 14% are correlated but discordant and 42% are not correlated [29]. These data led to the understanding that AIDs share a certain range of risk genes, which further points at specific immunological pathways [29, 43, 44].

1.4.3 Functional role of genetic variants

Although GWAS and Immunochip have given us a list of risk SNPs associated with various AIDs, the functional consequences of these risk variants remain elusive. Firstly, association signals often locate to large blocks of SNPs in strong LD, which makes it difficult to pinpoint the causal SNP [43]. Secondly, even though the causal risk SNP is evident, many risk SNPs are located in non-coding regions of the genome [23]. E.g. in RA, among SNPs in 100 non-MHC RA risk loci, only 20 % are missense or synonymous variants, whereas approximately 80% are non-coding (Figure 4).

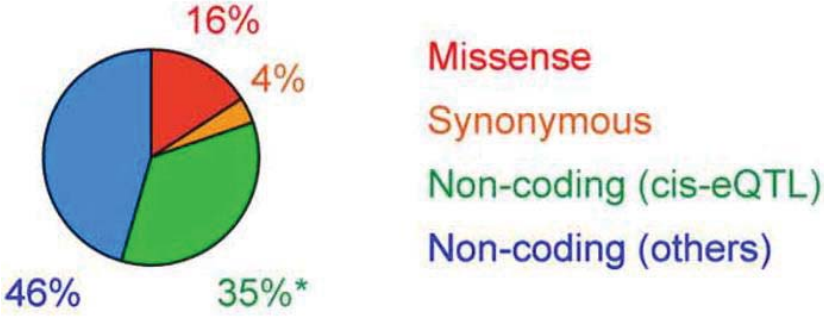

Figure 4: Functional annotation of SNPs in 100 non-MHC RA risk loci. In fact, 44% of all RA risk SNPs had cis-expression quantitative trait loci (eQTL), but 9 of them overlapped with missense or synonymous variants. 35% of them did not overlap as indicated by the asterisk. Figure from [45] with permission. eQTLs are described in more detail below.

The large number of non-coding variants has led to the suggestion that many risk SNPs may influence regulatory elements [46]. The next steps in deciphering how the AID risk variants contributes to the development of AID have therefore been to understand 1) which tissues and cell type the SNPs have an effect in, and 2) what types of regulatory element the SNPs interfere with.

1.4.3.1 Finding the tissues and cell types affected by AID risk variants

In 2011, Hu et al. sought to understand which tissues and cell types that are affected by the AID risk variants. They performed a study where 79 different human tissues were investigated, and because human cell types were not easily available at this time point, they also included 223 murine-sorted immune cells [47]. They found that GWAS loci associated with SLE, CD and RA were enriched with genes expressed mainly in human immune tissues, and moreover, in particular murine immune subsets (transitional B cells, epithelial-associated stimulated DCs and CD4+ effector memory T cells, respectively) [47]. Later studies have confirmed enrichment of genes in susceptibility loci that are preferentially expressed in human immune cell types, such as for instance in CD4+ T cell subsets [48].

1.4.3.2 Finding the regulatory elements affected by AID risk variants

Furthermore, many studies have tried to understand what kind of regulatory elements that could be affected by the AID risk variants. The activity of regulatory elements (such as promoters and enhancers) varies depending on the epigenetic landscape around the elements, and also by the presence and level of the corresponding transcriptional regulators binding to these sites. The epigenetic landscape is determined by a range of epigenetic factors such as DNA methylation and histone modification, as well as by chromatin remodelers and histone variants incorporated into the nucleosomes. The epigenetic landscape of each cell can vary considerably between loci and contributes to distinct gene expression programs and biological functions. Today, large effort has been put in deciphering the functional landscape in various human cell types and tissues. I will address two projects that have been actively involved in this process: the Encyclopedia of DNA Elements (ENCODE) project and the Roadmap Epigenomics project.

The ENCODE project was established in September 2003 by the National Institutes of Health (NIH) National Human Genome Research Institute (NHGRI). The goal of the ENCODE project was to identify all functional elements in the human genome sequence. This includes annotation of all the genes and their ribonucleic acid (RNA) transcripts (both protein-coding and non-coding), and all transcriptional regulatory elements [49]. To achieve these goals, NHGRI organized the ENCODE Consortium, an international collaboration of several research groups with expertise in producing and analyzing high-throughput functional genomic data. To facilitate comparison and integration of all the different data types, ENCODE have used selected sets of cell types. These include the widely studied EBV-immortalized B-lymphoblastoid cell line GM12878, the K562 erythroleukemia cells, the H1 human embryonic stem cell line (H1-hESC), HeLa-S3 cervical carcinoma cells, HepG2 hepatoblastoma cells, and primary human umbilical vein endothelial cells. Another set comprising more than 100 different cell types (https://genome.ucsc.edu/ENCODE/cellTypes.html), used to capture a broader spectrum of human biological diversity, is being used in selected assays. Today, the ENCODE project contains complementary genome-wide datasets including gene annotation (GENCODE), transcriptome analysis, chromatin structure and modifications (H3K27ac, H3K27me3, H3K36 me3 etc.), transcription factor binding sites, DNaseI footprints and DNA methylation. The Consortium releases data rapidly to the UCSC genome browser (http://genome.ucsc.edu/ENCODE/) and the ENCODE web portal (http://encodeproject.org).

A few years later, in 2007, the Roadmap Epigenomics Program was funded through the NIH Common Fund and established with the goal of exploring how epigenetics contribute to human health and disease [50]. As a first step, the Roadmap Reference Epigenome Mapping Consortium has generated a public resource (http://www.roadmapepigenomics.org) of genome-wide epigenetic maps in a broad range of human primary cells and tissues, frequently involved in human diseases. Both adult and fetal cells and tissues are represented, including stem cells, cells from a number of organs and distinct brain regions, and a variety of purified blood cell types (Figure 5). For each tissue and cell type analyzed, a dataset representing a “complete epigenome” is provided, including RNA expression, DNA methylation, a panel of histone modifications and DNase I hypersensitivity. The next step will be to apply these data to clarify the role of epigenetics in the development of complex diseases, and several investigators funded by the Roadmap Epigenomics Program are currently addressing these matter. The reference epigenomic maps cover a far broader range of human primary cells and tissues than was represented by ENCODE, and are immensely valuable to researchers trying to investigate how genetic variants contribute to disease.

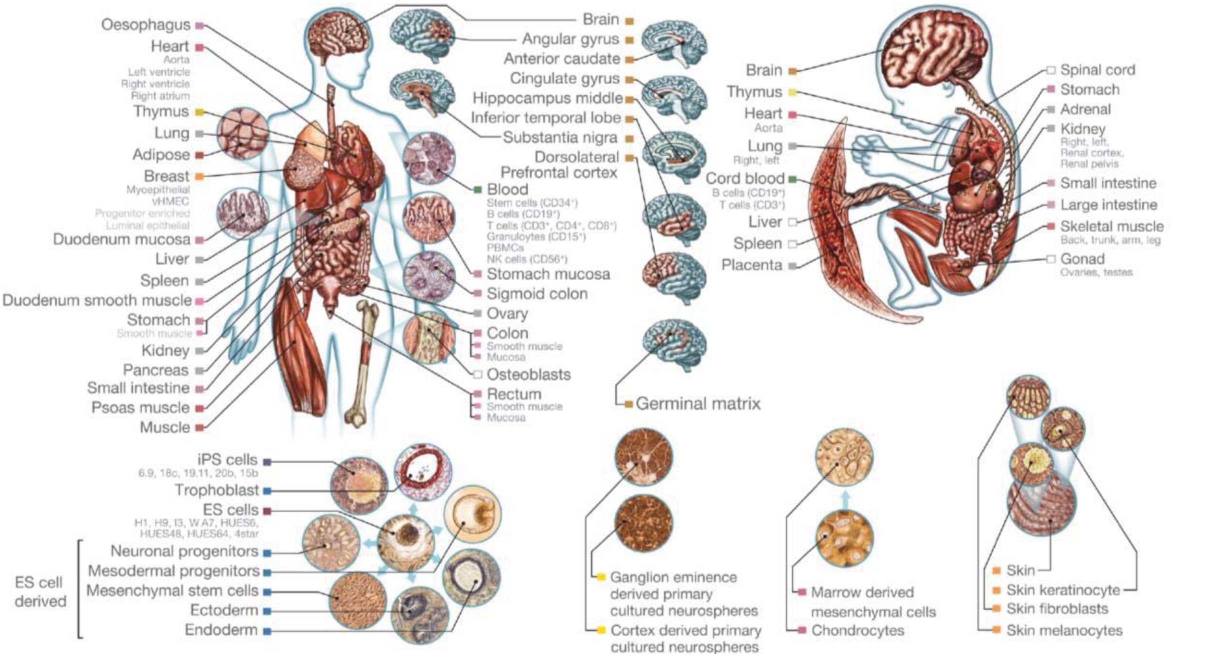

Figure 5: The cell and tissue types used in the Roadmap Epigenomics project. Figure from [51] with permission.

These projects have facilitated the functional prediction of variants from genetic association studies. For instance, one study in 2013 aimed at determining the localization of 2874 SNPs associated with 12 immune-mediated diseases, including both lead SNPs from GWAS studies performed in Caucasian individuals (n = 337) and their proxies (n =2537) in perfect LD (r^2 = 1) within a 500 kb region, relative to the gene structure in the genome [46]. They found that 2.6% of the SNPs mapped to exons, while 42.3% mapped to introns, 2.1% mapped to untranslated regions (UTRs) and finally 53% mapped to intergenic regions. When they further annotated these SNPs with regulatory sequence data from ENCODE, they found that 7.6%

mapped to promoter histone marks, 18.8% map to enhancer histone marks, 32.1% are in DNase-hypersensitive sites, 42.3% change motifs and 14.4% map to protein bound regions.

Furthermore, this study also found that multiple classes of regulatory, non-coding RNA molecules (including microRNAs and long intergenic non-coding RNA (lincRNAs)) was annotated to the AID risk variants [46]. Among the 2874 SNPs, 1.25% were reported to change microRNA binding sites and 8.52% mapped to lincRNAs [46]. LincRNAs is a type of long non-coding RNA (lncRNA) which, as specified by its name, is found in between coding genes. The definition of a lncRNA is "a non-coding RNA that have at least 200 nucleotides" and in humans, lncRNAs are often polyadenylated [52]. The molecular functions of lncRNA are highly diverse. LncRNAs can regulate a variety of processes that include transcription, splicing, RNA degradation and translation [52]. It has been reported that SNP variation can alter lincRNA [46, 53] expression levels and possibly affect downstream transcriptional programs.

Taken together, this clearly suggests that genetic variations that provide susceptibility to complex diseases modulate transcriptional regulatory mechanisms, which also means that these SNPs may affect the expression of nearby transcripts [46].

1.4.4 Expression quantitative trait loci

Many loci contain genetic variants that significantly correlate with gene expression; these are termed expression quantitative trait loci (eQTL). eQTL screening is practical for linking genetic variation to complex phenotypes [54]. When eQTL SNPs (eSNPs) and SNPs associated with complex diseases are located in the same region, this leads to a testable hypothesis that a genetic variant influence trait variance by affecting the expression of a given gene [54]. This is helpful in detecting potential causal genes for trait-associated variants, especially in regions where a large number of genes exist [54]. Usually, eQTLs are assessed by investigating SNPs and expression probes located up to 1 Mb apart [55]. These types of eQTLs are called cis-eQTLs. Cis-eQTL screens have been performed in many tissues, for instance in lymphoblastoid cell lines [56], liver [57], blood [58], brain [59], adipose tissue [60], skin [61] and primary fibroblasts [61]. Still today, both cis- and trans-eQTL screens are emerging rapidly. Interestingly, it has been reported that many eQTLs (71.3%) are concordant [55], meaning that a gene is regulated by a SNP in the same allelic direction with similar effect size in different tissues [55]. However, 28.7% of eQTLs are discordant and show tissue-dependent gene regulation [55], either by specific regulation, alternative regulation, different effect sizes or opposite allelic direction (Figure 6).

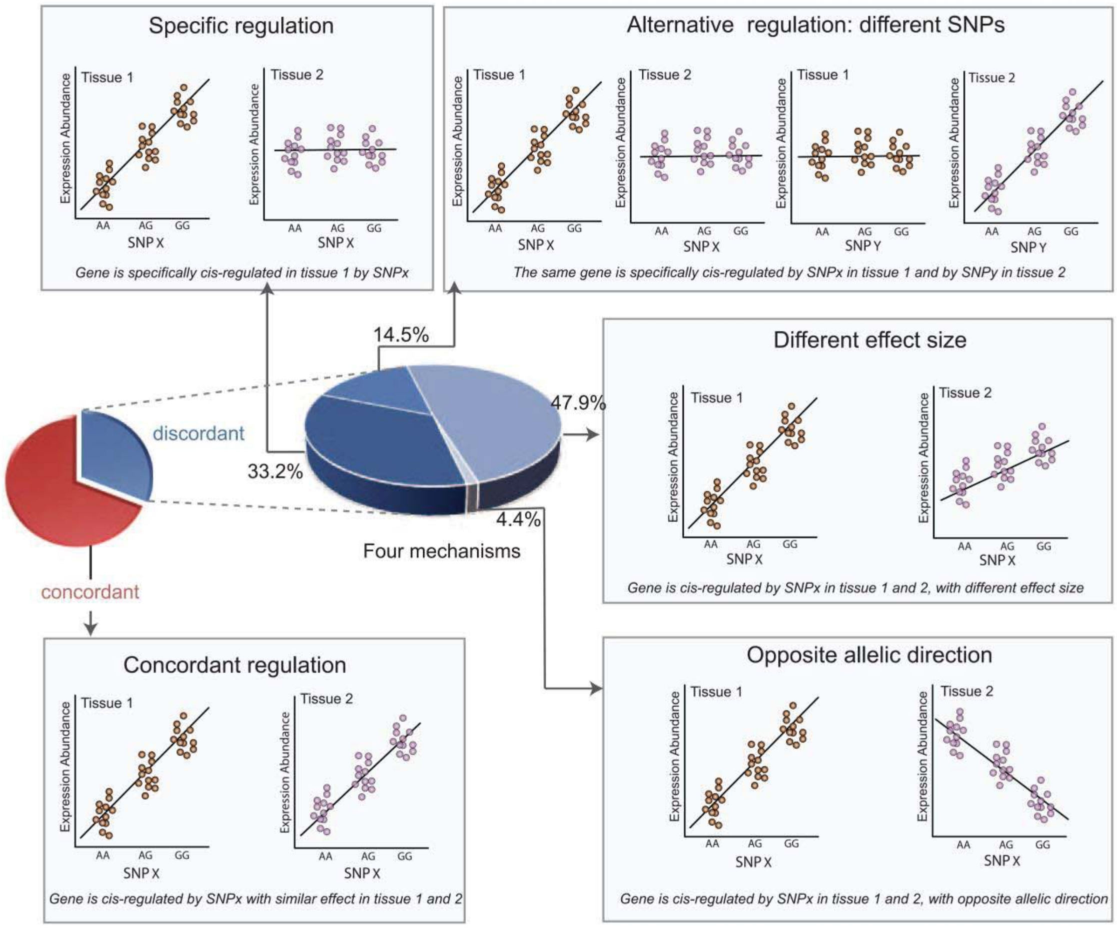
Specific regulation
Tissue 1
Tissue 2
Expression Abundance
AA AG GG
SNP X
Gene is specifically cis-regulated in tissue 1 by SNPx
Alternative regulation: different SNPs
SNP Y
The same gene is specifically cis-regulated by SNPx in tissue 1 and by SNPy in tissue 2
14.5%
47.9%
33.2%
4.4%
discordant
concordant
Four mechanisms
Different effect size
Gene is cis-regulated by SNPx in tissue 1 and 2, with different effect size
Concordant regulation
Gene is cis-regulated by SNPx with similar effect in tissue 1 and 2
Opposite allelic direction
Gene is cis-regulated by SNPx in tissue 1 and 2, with opposite allelic direction

Figure 6: Cis-regulation between tissues, figure from [55] (CC BY).

The NIH GTEX eQTL Browser (http://www.ncbi.nlm.nih.gov/gtex/GTEX2/gtex.cgi) and the Broad GTEx Portal database (http://www.gtexportal.org/home/) are public databases that provide eQTL screens performed in multiple tissues. Another public resource for SNP annotations that also offers eQTL information is the HaploReg V4 database.

HaploReg [62] (http://www.broadinstitute.org/mammals/haploreg/haploreg.php) is a comprehensive database which was launched in 2011. This database integrates SNP data with GWAS trait associations, eQTL data and functional annotation data from, among others, ENCODE and the Roadmap Epigenomics project and therefore enables the systematic interpretation of SNPs of interest. In the newest version of HaploReg (V4), chromatin state maps for 127 reference epigenomes are available from ENCODE 2012 and Roadmap

Epigenomics [63], and eQTL study results from the GTEx pilot analysis v6 and other eQTL projects are included.

Although large consortia have made tremendous efforts in mapping many types of human tissue today, not all tissues have yet been completely investigated. In this thesis, I have focused on the thymus organ, where the Roadmap Epigenomics project have established different epigenetic maps, however, only three eQTL studies [64-66] have been performed in this organ, assessing a small number of AID risk variants (n = 52). Furthermore, high throughput RNA sequencing of central thymic cell types, like for instance certain thymic antigen presenting cell (APC) types, have to date not been performed. Therefore I will now describe this organ in more detail.

1.5 The thymus

The thymus is a specialized primary lymphoid organ of the immune system, and the site for proliferation and maturation of the T cell precursors, namely the thymocytes. This organ is essential for the establishment of "central tolerance", i.e. the generation of a repertoire with functional and self-tolerant T cells. To achieve this, the developing thymocytes need to pass several control points in the thymus involving several highly specialized thymic stromal cells. A large fraction (90-95%) of all thymocytes is lost due to death in the cortex [67-69], and another 50-70% of the positively selected thymocytes is thought to be subject to negative selection in the medulla [67, 70]. This reflects the strictly regulated selection process of self-tolerant T cells that are let out in the periphery. Below I described the journey of a thymocyte through the thymus, and further describe in more detail four highly specialized thymic cells implicated in the selection process of thymocytes to avoid autoimmunity.

1.5.1 Thymocytes develop and go through several selection points in thymus

The thymus is composed of two identical lobes, which can be further divided into several small lobules. Each lobule has an outer capsule, a peripheral cortex and an inner medulla.

Progenitor cells from the bone marrow enter the thymus through the cortex (Figure 7). After commitment to the T cell lineage, the thymocytes rearrange their TCR genes and become either γ δ or α β thymocytes at the double negative (DN) stage [71]. A fraction of the α β DN cells further gives rise to a large number of CD4 and CD8 double positive (DP) thymocytes. The TCR genes undergo somatic rearrangement, which results in a diverse repertoire of distinct α β TCRs with different specificities. Inside the cortex, the DP thymocytes will encounter the first type of specialized APC, namely the cortical thymic epithelial cell (cTEC).

cTECs present self-peptides to the DP thymocytes by their major histocompatibility complex (MHC) molecules. The DP thymocytes with TCRs that do not recognize a self-peptide-MHC complex will die by neglect, whereas those with TCRs with intermediate affinity and/or avidity for self-peptide-MHC complex will receive a positive selection signal and differentiate into single positive (SP) CD4+ or CD8+ cells [72].

SP thymocytes progress further into the medulla. The thymic medulla is the site for T cell tolerance induction. Any interference with the thymic medulla will manifest in autoimmunity, whether it is disruption of the three-dimensional space, disturbance of the development of the medullary cells, disrupted transition of SP thymocytes from the cortex or premature departure of thymocytes from the medulla [72].

The SP thymocytes encounter at least four different thymic medullary APCs; the medullary thymic epithelial cells (mTECs), the cluster of differentiation (CD)141+ and CD123+ DCs and finally, thymic B cells. SP thymocytes with TCRs that bind with too high affinity to the self-peptides presented by the medullary APCs are considered potentially harmful to the individual, and will be deleted (clonal deletion). Alternatively, SP thymocytes that bind with intermediate affinity may differentiate into regulatory T cells (Tregs). The thymocytes with low affinity to the self-peptide MHC complex are accepted as tolerant, and will be released into the periphery.

Nevertheless, a small fraction of autoreactive T cells are present in the blood of healthy individuals [10], indicating that escape from the strict control mechanisms in the thymus do occur. Mechanisms to ensure self-tolerance exist in the periphery as well, such as clonal diversion (selection of Tregs), receptor editing and anergy [71], but will not be further discussed here.

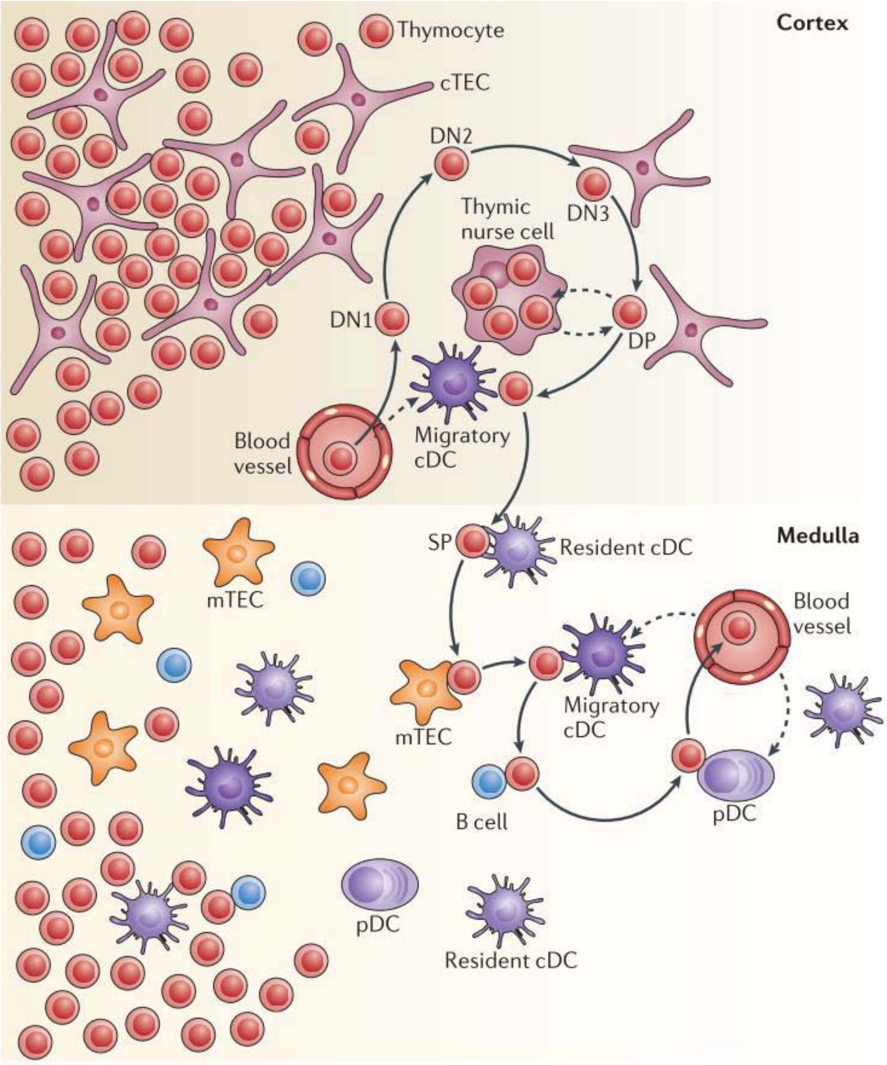

Figure 7: The journey of a developing thymocyte in thymus. Figure from [72] with permission.

1.5.2 Thymic antigen presenting cells in the medulla

1.5.2.1 medullary thymic epithelial cells

The mTECs are specialized epithelial cells in the thymic medulla that can transcribe a broad range of tissue-specific genes [73]. The expression of genes encoding tissue-restricted antigens (TRA) in mTECs contrasts with the tight spatio-temporal control of gene expression in peripheral tissues during pre- and post-natal development and is referred to as "promiscuous gene expression" [74]. Although many peripheral tissues are presented, a given TRA is only expressed in a minority of mTECs (1-3%) at any given time [72]. Also, groups of 100 - 300 TRA genes are usually co-expressed in subsets of human mTECs [75]. These groups of TRAs are often localized together within nuclear subdomains.

Around 40% of the TRA genes [76] are expressed under the transcriptional control of the autoimmune regulator (AIRE) protein. AIRE is a transcription factor crucial for the establishment of central tolerance, and loss-of-function mutations in *AIRE* in humans cause a recessive autoimmune syndrome termed autoimmune polyendocrinopathy-candidiasis-ectodermal dystrophy (OMIM: 607358). Although the exact mechanism for how AIRE regulates the expression of TRA genes remains elusive, AIRE is believed to recognize histone H3 tails hypomethylated at Lys4 (H3K4me0) [77]. This is a histone modification typically found in transcriptionally inactive chromatin, such as in the silenced tissue-specific gene loci in mTECs. Furthermore, AIRE interacts with P-TEFb [78] and other transcriptional regulators (reviewed in [79]) to promote gene expression through the release of stalled RNA polymerase [80] and elongation of RNA transcripts.

The mTECs express and present TRAs in context of both MHC class I and MHC class II molecules [67]. In terms of MHC class I presentation, mTECs and their fellow medullary APCs express both the housekeeping proteasome (characterized by its β5-unit, encoded by the *PSMB5* gene) and the immunoproteasome (containing a β5i-unit, encoded by the *PSMB8* gene). For MHC class II presentation, mTECs uses a distinct pathway. Through autophagy ("self-eating"), endogenous proteins from the cytoplasm are sequestered into double-membrane-delimited compartments (termed autophagosomes) and delivered to lysosomes responsible for MHC class II loading. There, proteases, such as cathepsin S [81], process the substrates before MHC class II loading and presentation on the cell surface.

1.5.2.2 Dendritic cells

mTECs are efficient in inducing clonal deletion of thymocytes that bind the MHC-peptide complex with high affinity. TRAs expressed in mTECs can also be transferred to and cross-presented by thymic DCs [81].

DCs make up 0.5% of the cells in thymus, and are important mediators of central tolerance. Both conventional DCs (cDCs) and plasmacytoid DCs (pDCs) are found in thymus, where the cDCs constitute around two-thirds of the thymic DC population. The major subset of cDCs (roughly 2/3) can further be classified as CD8α+SIRPα- cDCs in mice [72] or CD141+ in humans [82]. The CD8+ SIRPα- cDCs are commonly referred to as the resident cDCs, as they arise from an intrathymic differentiation pathway [72]. There is also accruing evidence that these cDCs can present TRAs that have been transferred by mTECs, although the direct handover has been technically challenging to prove [72]. The minor subset of cDCs in thymus (roughly 1/3) can be characterized as CD8- SIRPα+ cDCs in mice and CD11b+ in humans [82, 83]. These are migratory cDCs which are enriched in cortico-medullary perivascular space and also around small vessels, where they effectively capture and present blood borne antigens [84]. Finally, the last DC type present in both mouse and human thymus are the pDCs (characterized as CD123+ in humans [82]). These are also migratory DCs, which capture and transport peptides from the blood stream into the thymus, where they present the peptides to the SP thymocytes [85]. DCs of thymic origin have been reported to express an endogenous Mtv-encoded superantigen to developing thymocytes to induce negative selection [86]. However, we do not know to which extent DCs are capable of transcribing TRAs naturally in human thymus.

1.5.2.3 Thymic B cells

Approximately 0.3% of the cells in thymus are B cells [72]. Thymic B cells are also capable of presenting peptides to the developing thymocytes and induce negative selection [87-90]. The origin of thymic B cells is not fully understood, as development from intrathymic progenitors [87] and migration from the peripheral circulation [89] have been suggested. Until 2015, it was believed that thymic B cells were only capable of presenting peptides derived from antigens captured through their BCR. However, Yamano et al. then discovered a small population of non-epithelial, Aire positive cells, the majority of which also expressed the B cell marker CD19 [89]. This subset of thymic B cells was found to transcribe Aire and Aire-dependent TRAs [89, 91]. Recently, Gies et al. performed high-throughput RNA sequencing and confirmed expression of *AIRE* and a few TRA genes in a subset (5%) of human thymic B cells [92]. These findings give us novel insight into the functional role of B cells in thymus.

Taken together, the thymus, and in particular the thymic APCs, have an important role in protecting the body against autoimmunity. Furthermore, as risk variants associated with AIDs have regulatory roles in immune related cell types, and because trait-associated SNPs from GWAS are more likely to be eQTLs [93], it is important to further investigate whether AID risk variants can influence the transcriptional landscape in the thymus organ. Now that the most important topics have been covered in the introduction, I will further describe the aims of this thesis.

2. Aims

The overall aim of this thesis was to investigate whether AID risk variants influence thymic gene expression.

The following objectives were addressed:

- To assess whether fine-mapped AID risk variants influence the thymic gene expression of *ERAP1* and *ERAP2*, encoding two aminopeptidases important for peptide processing before loading onto HLA-class I molecules (**paper I**)
- To systematically investigate whether gene expression levels in thymus could be influenced by AID risk variants (**paper II**)
- To characterize the transcriptomes of four thymic APCs (**paper III**)
- To investigate whether genes being thymic eQTLs for AID (**paper I** and **II**) were expressed in the thymic APCs (**paper III**).

3. Summary of papers

3.1 Paper I

Autoimmune risk variants in ERAP2 are associated with gene expression in thymus

The endoplasmic reticulum aminopeptidases ERAP1 and ERAP2 cleave peptides before loading onto the MHC class I molecules and presentation on the cell surface. Multiple genetic risk variants in *ERAP1* and *ERAP2* associated with different AIDs (AS [94, 95], CD [96], IBD [36], MS [97], Ps [98] and T1D [99]) have been identified through GWAS. GWAS risk variants have further been fine-mapped for several AIDs. Since no eQTL screen with *ERAP1* and *ERAP2* has yet been performed in thymic tissue, we performed an eQTL screen to investigate whether seven fine-mapped AID SNPs in the *ERAP*-region influence the gene expression levels of *ERAP1* and *ERAP2* in thymus.

This study correlated genotypes of seven fine-mapped AID SNPs with gene expression levels from all probes within a window of +/- 1 Mb in the *ERAP*-region. The genotypes were obtained by Immunochip and gene expression levels from human thymic tissues (n = 42) were measured on an Illumina HumanWG-6 v3 microarray. After quality control, six significant SNP-probe pairs were evident, involving two eQTL probes (eProbes) binding to *ERAP*1 and *ERAP2*, respectively. When the eProbes were further tested against all SNPs on the Immunochip densely covering the *ERAP*-region, two independent peak eQTL signals were detected in *ERAP1* and *ERAP2*, respectively. Interestingly, the peak eQTL signal overlapped with the AID risk loci in *ERAP2* ($r^2 > 0.94$), but were distinct in *ERAP1* ($r^2 < 0.4$). We discovered that among the SNPs showing the most significant eQTL associations with *ERAP2* ($P < 3.4 \times 10^{-20}$), six were located in transcription factor motifs within an open chromatin region that had epigenetic histone marks suggestive of promoter (H3K4m3) and active enhancer (H3K27ac) function in human thymus. We also observed an association between the haplotype comprising all the risk alleles and the highest level of *ERAP2* expression.

This study therefore reveals highly correlated, fine-mapped AID risk variants that act as eQTLs with *ERAP2* in thymus, and further highlights potential causal regulatory variants. The most significant eQTL in *ERAP1* was in low LD with the AID risk variants, and the eQTL signals from the AID SNPs were markedly inferior. Hence, the thymic eQTL involving *ERAP1* is distinct from the *ERAP1* associations observed in several AIDs. These findings show the importance of thoroughly mapping eQTL signals in tissues.

3.2 Paper II

Genetic variants for autoimmune diseases that influence gene expression in thymus

GWAS have enabled the identification of hundreds of genetic risk variants for AIDs. The majority of these risk variants are located in non-coding regions, thus indicating that they may have a role in gene regulation. A common approach to connect genetic variants with the pattern of gene expression differences is by performing eQTL analysis. Before the initiation of this project, most eQTL studies had been executed in blood, and only three studies [64-66], performed in our research group, had addressed the influence of risk variants on gene expression levels in thymus. The thymus plays an essential role in the development of immune tolerance, as it is the organ where the T-lymphocytes are selected through positive and negative selection to avoid the release of autoreactive T cells into the periphery. Failure in the immune self-tolerance system is a hallmark of AIDs. In the previous thymic eQTL screen, CeD associated risk variants were significantly ($P < 0.05$) associated with gene expression levels in thymus.

This study, therefore, sets out to determine whether 353 GWAS identified risk variants associated with 11 different AIDs influence gene expression in thymus. Genotypes of the 353 GWAS risk variants were obtained by Immunochip and tested against expression levels, measured by the Illumina HumanWG-6 v3 microarray, of surrounding genes (+/- 1 Mb) in human thymic tissue (n = 42). Using a stringent significance level, we identified eight eQTLs located within seven genetic regions (*FCRL3*, *RNASET2*, *C2orf74*, *NPIPB8*, *SIRPG*, *SYS1* and *AJ006998.2*) where the expression was associated with AID risk SNPs at a study-wide level of significance ($P < 2.7 \times 10^{-5}$). In *NPIPB8* and *AJ006998*.2, the eQTL signals appeared to be thymus-specific. Furthermore, since GWAS risk variants had later been replicated and fine-mapped in several studies employing the Immunochip, we searched for fine-mapped AID SNPs (+/- 1 Mb) located within each of the eQTL gene regions. Fine-mapped AID SNPs were in strong LD ($r^2 > 0.8$) with the thymic eSNPs within *RNASET2* and *SIRPG*. These fine-mapped AID SNPs were also associated with the same diseases as our GWAS selected eSNPs. Finally, in all eQTL regions, except *C2orf74*, SNPs underlying the thymic eQTLs were predicted to interfere with transcription factors important in T cell development.

Our study, therefore, provides evidence for autoimmune risk variants that act as eQTLs in thymus, and suggest that thymic gene regulation may play a functional role at some AID risk loci.

3.3 Paper III

Transcriptomes of antigen presenting cells in human thymus

Thymic APCs play a crucial role in establishing a repertoire of functional and self-tolerant T cells to prevent autoimmunity. To date, most studies concerning thymic APCs have been performed in mice. With the exception of thymic B cells, no one has yet explored the transcriptomes of human thymic APCs. Therefore, we performed high throughput RNA sequencing of four primary APCs in human thymic tissue to compare their transcriptomes and investigate specific genes important for APC function.

We isolated six biological replicates of mTECs, CD19+ B cells, CD141+ and CD123+ DCs from human thymic samples. One biological replicate was removed from the mTECs due to contamination. We performed high throughput RNA sequencing and used the EdgeR software to compare their transcriptomes. We found that thymic CD141+ DCs and mTEC expressed the highest and the lowest levels of all classical HLA genes, respectively. Among 21 genes encoding proteins involved in the HLA class I and II pathway, we found that 14 (67%) had the highest gene expression levels in CD141+ DCs and consistently the lowest levels in mTECs. Three transcriptional regulators of TRA expression were also investigated, where *DEAF1*expression was detected in all four APCs but the highest levels were clearly found in CD141+ DCs. Expression of *AIRE* and *FEZF2* was mainly found in primary human mTECs. We further investigated how the repertoire of "tissue enriched genes" from the Human Protein Atlas was distributed between the thymic APCs. These genes were defined as "TRA genes" in our study. We detected expressed (Fragments per Kilobase of transcript per Million mapped (FPKM) reads > 1) TRA genes in all four APCs, but the mTECs were clearly dominating in both the total number and in the number of unique genes expressed. The percentage of uniquely expressed TRA genes (FPKM > 1) was 20% in mTECs, 7% in CD19+ B cells, 4% in CD123+ DCs and 2% in CD141+ DCs. In the mTECs and B cells, some of the unique genes encoding TRAs also overlapped with reported human autoantigens from the literature. Finally, we investigated whether any of the previously reported thymic eQTLs associated with AID ("AID-eQTLs") were uniquely expressed in any of the thymic APCs. Expression of the eGenes underlying the thymic AID eQTLs were indeed detected in several thymic APCs, however, none were uniquely expressed. This study offers, to our knowledge, the first transcriptome data of mTECs, CD141+ and CD123+ DCs isolated from human thymic tissue and provides an overview of particular genes important in APC function that varies significantly between the cell types.

4. Methodological considerations

4.1 Study population

The thymic tissue samples used in this thesis were obtained from children under the age of 13, undergoing cardiac surgery. A total of 42 thymic tissues were included in **paper I** and **II**, and 6 separate tissues were included in **paper III**. In **paper I** and **II**, the gender distribution was 22 girls and 20 boys. Because gender can be a confounding factor, we did not include any risk variants from the GWAS catalog located on the X and the Y chromosomes. Furthermore, the age range varied from 4 days to approximately 13 years. As age is a second possible confounder, we investigated whether the gene expression levels (and hence the eQTLs) could potentially be influenced by the age variation (Figure 8). However, we could not detect any specific correlation between age and gene expression for the eQTL genes.

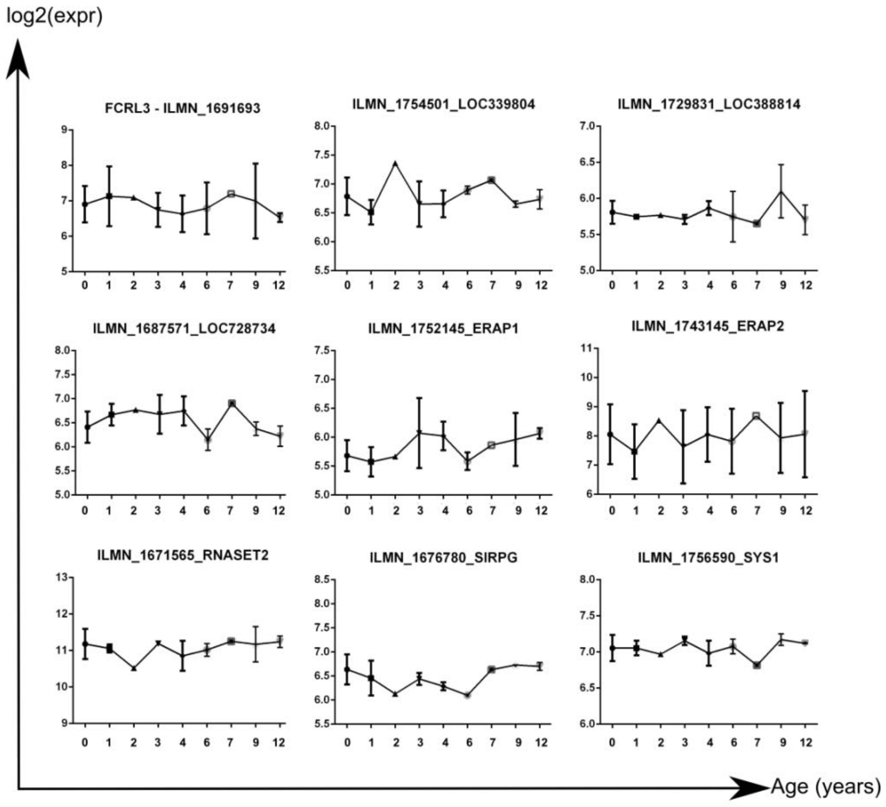

Figure 8: Gene expression levels (log2) of the 9 eGenes underlying the thymic AID eQTLs in paper I and II as a function of age

In **paper III**, only boys were included (n = 6), which means that there are boy-specific differences between the APCs in our dataset, such as expressed genes located on the Y chromosome that are not in the pseudoautosomal regions. Moreover, the age range varied from 5 days to 1 year and 4 months. Because we are comparing cell types and not individuals, the “age average” (n = 271 days) is equal in the compared groups. The exception is the mTECs, where one biological replicate was removed due to contamination by DCs. This replicate was 334 days old, and therefore a slightly different “age average” (n = 257 days) is found in this cell population, which possibly could have influenced the expression levels in our dataset.

In **all three papers**, we only included patients without any known syndromes or AIDs, as genetic effects on a phenotype can be confounded by the presence of a disease [100]. All patients were of Norwegian origin, as ethnical differences can influence allele frequencies in genetic analyses. To our knowledge, cardiac diseases or surgery does not affect the gene expression profile in thymus. However, this cannot be completely disregarded. Informed consent has been given from the parents of the children undergoing heart surgery to use the thymic tissue for research purposes. Ethical approval was obtained by the Regional Committee for Research Ethics (REK approval number: S-04101).

4.2 Thymic tissue preparation

4.2.1 Cell separation

In **paper I** and **II**, whole thymic tissue was used for DNA and RNA extraction. However, in **paper III**, we first isolated four individual APCs from human thymus: mTECs, CD141+ DCs, CD123+ plasmacytoid DCs and CD19+ B-cells. The first years of my PhD was used to establish a comprehensive protocol to isolate these APCs based on the work of others [101-105]. The isolation procedure has been described in detail in **paper III** and is illustrated in Figure 9. The aim of our study was to analyse and compare the transcriptomes of primary APCs; therefore we isolated the cell populations the same day as the thymectomy. We received about 10 g of thymic tissue every time (which corresponds to about half a thymus of a child < 1 year).

The thymus is composed of 98% thymocytes and only 2% APCs [102], however, cell culture of the APCs was not an option for us, as cell culture has shown to alter the transcriptional profile of other human primary cell types [106]. We wanted an accurate representation of the APC transcriptomes, as close as possible to their in vivo representation. In large-scale differential expression RNA sequencing studies, it has previously been reported that adding biological replicates increases power to detect differentially expressed genes [107]. We therefore chose to isolate six biological replicates for each APC type.

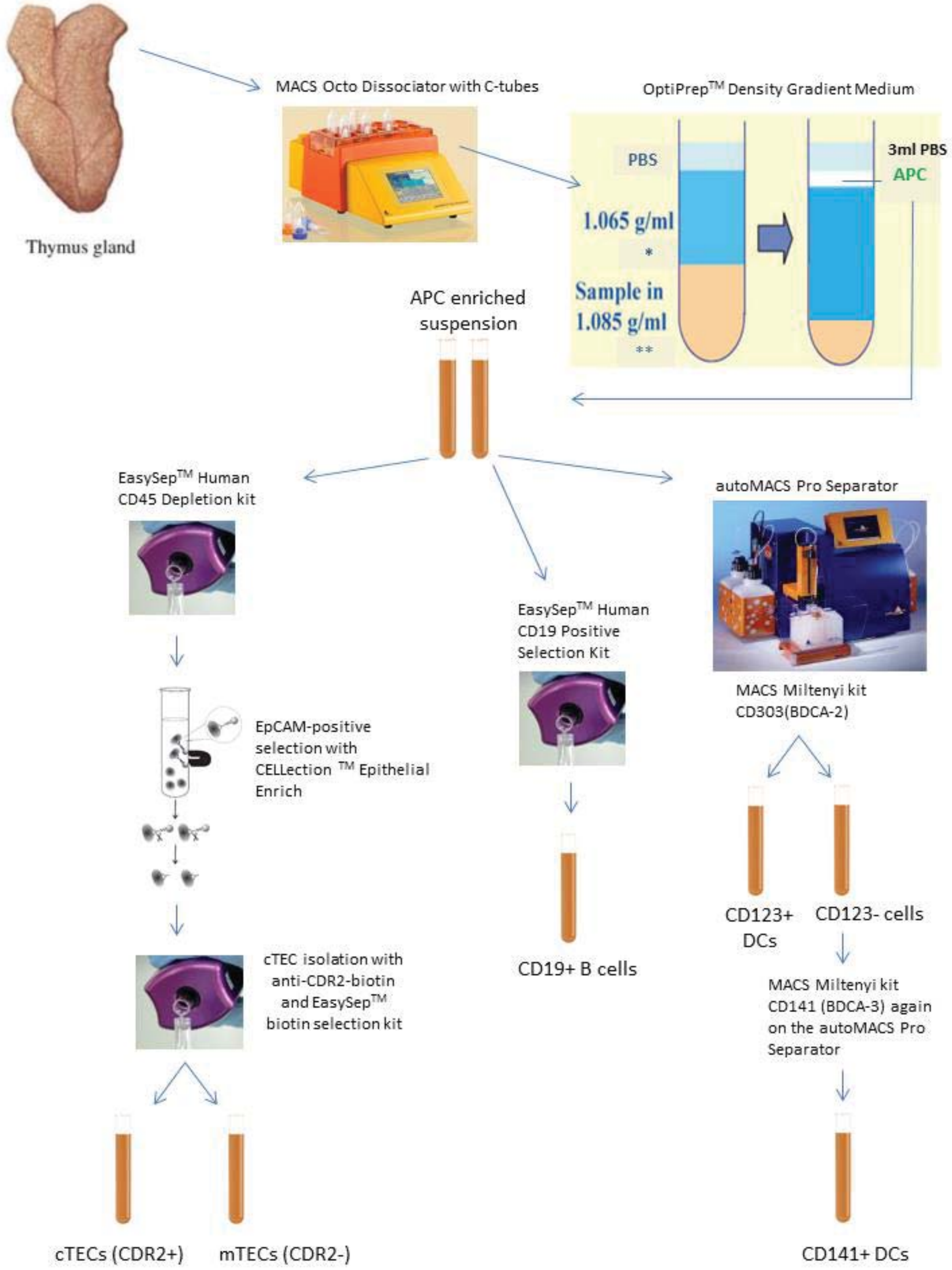
Thymus gland
MACS Octo Dissociator with C-tubes
OptiPrep™ Density Gradient Medium
PBS
1.065 g/ml
*
Sample in
1.085 g/ml
**
3ml PBS
APC
APC enriched suspension
EasySep™ Human CD45 Depletion kit
EpCAM-positive selection with CELLection ™ Epithelial Enrich
cTEC isolation with anti-CDR2-biotin and EasySep™ biotin selection kit
cTECs (CDR2+)
mTECs (CDR2-)
EasySep™ Human CD19 Positive Selection Kit
CD19+ B cells
autoMACS Pro Separator
MACS Miltenyi kit CD303(BDCA-2)
CD123+ DCs
CD123- cells
MACS Miltenyi kit CD141 (BDCA-3) again on the autoMACS Pro Separator
CD141+ DCs

Figure 9: Procedure of thymic APC purification. *An 11.5% iodixanol solution (ρ = 1.065 g/ml) was made with OptiPrep^{TM} and a diluent consistent of Phosphate-Buffered Saline (PBS), Fetal Bovine Serum (FBS) and Ethylenediaminetetraacetic acid (EDTA).** The cells were suspended in PBS and OptiPrep^{TM} to make a 15% iodixanol solution (ρ = 1.085 g/ml).

We started preparing the tissue by removing fat and necrotic material. Furthermore, because the thymus is composed of only 2% APCs [102], a combination of gentle enzymatic digestion steps and enrichment protocols is required to efficiently isolate the APCs with minimal cellular damage and maximal yield [104]. We treated our thymic tissue with two enzymatic solutions, first with Collagenase D (Roche Life Science, Switzerland) and then with Liberase^{TM} TM (Roche Life Science) [104] in two C-tubes on a gentleMACS Octo Dissociator (Miltenyi Biotec, Germany). In this way, we always managed to completely dissolve the entire tissue. Liberase research grade enzymes are in fact a blend of highly purified collagenases [104], and the use of collagenases instead of a highly digestive enzyme such as trypsinase results in preservation of cell surface markers [103].

Although the recommended incubation temperature for both Collagenase D and Liberase^{TM} TM is 37°C according to the manufacturer, it has been reported that incubation with Collagenase D in room temperature is a more gentle way of handling DCs in order to not activate them [101]. We postulated that this treatment would be better for the DC and B cell surface markers (CD123, CD141 and CD19), and therefore treated the first C-tube intended for DC and B cell isolation at room temperature (~20°C). The second C-tube, intended for TEC isolation, was treated at 37°C [102]. We observed high cell viability when counting the cells after both 20°C and 37°C treatment; however, whether this difference in temperature might have had an implication on the RNA profile of the APCs is hard to say. As the normal body temperature is 37°C, genes encoding heat shock proteins are not activated. If anything, the cells in the 37°C tube could potentially be slightly more stressed compared to the cell in the 20°C tube due to the higher optimal enzyme activity.

We further separated the APCs from the thymocytes based on the cell density gradient medium OptiPrep^{TM} (Axis Shield, Norway). OptiPrep^{TM} is a 60% (w/v) solution of iodixanol in water and is used to separate light density cells from tissues such as Peyer's patches, blood, lymph nodes, spleen, thymus and Langerhans cells from skin. The fraction of light density cells in thymus is highly enriched in APCs, both DCs and TECs [102]. The clarity of the band varied from time to time, indicating that we had variable amounts of APCs from the different thymic tissue pieces.

The "APC bands" (one treated at 20°C and one at 37°C) were further transferred to two new respective tubes, before we used magnetic microbead kits to isolate the respective APC types. The kits from MACS Miltenyi magnetically label the DCs with antibodies directed against the specific marker (CD141 or CD303) either directly (CD141), or indirectly through antibody-biotin and anti-biotin-MicroBead complexes (CD303).

The kit from EasySep^{TM} uses dextran-coated magnetic nanoparticles in addition to bispecific tetrameric antibody complexes directed against both dextran and the CD19 surface marker on B cells, or the CD45 marker on hematopoietic cells. The latter was used to deplete the tube intended for TEC isolation before positive selection with Dynabeads that were coated with the monoclonal mouse IgG1 antibody Ber-EP4 (Anti-EpCAM). In this way, fibroblasts and endothelial cells that are EpCAM- were removed from the TEC suspension. The cTECs (EpCAM+CDR2+) were separated from mTECs (EpCAM+CDR2-) by anti-CDR2-biotin and EasySep^{TM} biotin positive selection kit.

All kits, except the Human CD19 Positive Selection Kit, reported their feasible use on other tissues than peripheral blood, as long as a single-cell suspension was made by standard preparation methods. However, complex tissues such as thymus have vast cell heterogeneity, and all thymic cell types may not have been mapped to date. Using antibodies in these kinds of tissues increases the risk of including cell types that express the same markers as our APCs of interest. We might therefore have included yet-unidentified thymic cell types in our samples (**paper III**).

At the end of each isolation, we counted the number of cells we obtained (Table 2) on a Countess™ Automated Cell Counter (Thermo Fisher Scientific, USA).

Table 2: Median cell yields and interquartile range (IQR) from the APC samples.

APC	**Median number of isolated cells (IQR)**
CD123+ DC	32,000 (24,250)
CD141+ DC	47,500 (47,000)
CD19+ B cells	90,000 (97,500)
mTEC	174,000 (147,000)

The final cell populations were stored in RNAprotect® Cell (Qiagen, Germany) after ended isolation procedure, which provides immediate stabilization of RNA and thus the gene expression profile in the APCs. This usually occurred within 7-8 hours after the thymus had been removed from the patient by surgery. The length of the protocol and all the kits used in this procedure might have affected the cell viability and the RNA transcriptome profile of our thymic cell types (**paper III**). The mTEC isolation was particularly long, as they had to go through three different processes (CD45 depletion, EpCAM positive selection and finally cTEC removal) after the APC band was collected from the centrifuged OptiPrep^{TM} solution, as compared to the CD123+ and the CD19+ cells (ready after one kit) or the CD141+ cells (ready after two kits).

A more suitable method for isolating individual thymic cell populations would have been by cell sorting on a FACSAria flow cytometer (BD Biosciences). This technology could have shortened the isolation procedure with two-three hours, but was unfortunately not accessible at our department.

4.2.3 DNA and RNA extraction

In **paper I and II**, DNA and RNA were extracted from whole thymus by using TRIzol® reagents (Thermo Fisher Scientific). TRIzol enables the simultaneous isolation of RNA, DNA and proteins from biological materials [108]. In **paper III**, RNA was extracted from all cell types with RNeasy Plus Micro Kit (Qiagen, Germany), which purifies all RNA molecules over 200 nucleotides. This procedure enriches for messenger RNA (mRNA) and lncRNA, as most RNA molecules less than 200 nucleotides (such as 5.8S rRNA, 5S rRNA, and tRNAs) are selectively excluded. RNA concentrations and the RNA integrity numbers (RIN) were measured on a Nanodrop spectrophotometer (Thermo Fisher Scientific) and on a Bioanalyzer 2100 instrument (Agilent Technologies, USA). From the 42 thymic samples (**paper I and II**), we obtained a median RNA yield of 53 ± 24 µg and a median RIN of 7.9 ± 0.7. For the thymic APCs (**paper III**), the median RNA yields and RIN are showed in Table 3. An unexpected finding was that the mTEC RNA concentrations turned out to be the lowest, especially when the median cell number was higher for mTECs compared to the other APCs. It has recently been reported that the TRA repertoire in mTECs are overlapping between the individual developmental stages [109], suggesting that our finding is not because mTECs vary in the amount of TRA transcripts they produce during the different phases of maturation. However, one possible explanation for the low RNA yields could be, as mentioned above, the long isolation time for these cell types. The RNA concentrations in the mTEC samples were unfortunately too low to measure the RIN, restricting the possibility to investigate whether the RNA was degraded in these samples.

Table 3: Median RNA yields, RIN and IQR from the APC samples after extraction with RNeasy Plus Micro Kit (Qiagen). The mTEC concentrations were too low to measure the RIN.

APC	Median RNA yield (IQR)	Median RIN (IQR)
CD123+ DC	2.6 ng (1.32)	9.1 (1.1)
CD141+ DC	6.3 ng (14.08)	9.35 (0.52)
CD19+ B cells	2.6 ng (2.45)	9.5 (0.3)
mTEC	0.13 ng (0.11)	- (-)

4.3 Selection and genotyping of AID risk variants

In **paper I**, we searched for risk variants ($P < 5 \times 10^{-8}$) associated with AS [31], ATD [33], IBD [36], JIA [41], MS [37], Ps [34], RA [32] and T1D [42] on the Immunochip in the ERAP-region (defined as chr5: 95974244-96474244 where coordinates are based on the Genome Reference Consortium Human Build 37 (GRCh37) or Hg19). The seven selected SNPs were firmly associated with AIDs with P-values $< 7 \times 10^{-9}$. In **paper II**, risk variants associated with AS [42, 94], CD [96, 110, 111], CeD [35, 111, 112], IBD [113, 114], MS [115-127], Ps [98, 128-130], RA [131], SLE [132-135], T1D [136-138] and UC [139] were selected from the GWAS catalog. The last time we accessed the catalog was January 2012. Although the genome-wide significance threshold is $P < 5 \times 10^{-8}$, the GWAS catalog also includes individual SNP-trait associations identified in eligible studies with P-values $< 1.0 \times 10^{-5}$. The 393 selected risk SNPs had AID associations with P-value $< 1.0 \times 10^{-6}$. In both **paper I** and **II**, we only chose SNPs from AID association studies performed in Caucasians since the thymic tissue originated from Norwegian patients. In **paper I** and **II**, genotyping of the 42 Norwegian thymic tissues was performed on the Illumina Immunochip v1 (Illumina, San Diego, CA, USA) array at a core facility in Kiel (http://www.ikmb.uni-kiel.de/resources/genotyping). The Immunochip is a SNP genotyping array with dense marker coverage across 186 genetic regions. The probes on this array cover 195,806 SNPs and 718 small insertions-deletions [29]. The Immunochip has successfully enabled researchers to replicate previous GWAS findings, and also to fine map the peak association signal.

However, one limitation of the Immunochip is that it was designed using early 1000 Genomes Pilot data [29] with only 180 samples. Therefore, the coverage of the Immunochip is not complete. Secondly, approximately 10% of SNPs failed the assay design and have therefore not been included on the array [29]. Finally, as the Immunochip used was designed in 2009, AID associated risk SNPs identified through GWAS after 2009 might not be covered by the

Immunochip. We noticed these limitations particularly for the locus harboring the *FCRL3* gene (**paper II**), where few SNPs (n = 9) were available for the eQTL screen, and where novel AID associations have emerged later (for instance, rs2317230 associated with RA [45] and located in close vicinity to the *FCRL3* probe (28496 bp) was not present on the Immunochip). For other loci, the limitations of the Immunochip indicate that we might have lost valuable SNP information.

4.4 Gene expression measurement

4.4.1 Microarray

In **paper I** and **II**, gene expression levels in the 42 thymic tissues were measured at the Norwegian Genomics Consortium using the HumanWG-6 v3 microarray (Illumina, San Diego, CA, USA), comprising 48802 probes. The use of microarrays facilitates the quantitative measurement of thousands of genes simultaneously. The microarray contains multiple microscopic DNA spots, and each spot encompass specific probes for cDNA hybridization. Microarrays are suitable for screening high numbers of samples in a relatively cost and time efficient manner, hence this was considered to be a good choice at the time.

However, there are a few disadvantages with microarrays in regard to whole genome expression profiling. One limitation is high background levels due to cross-hybridization [140]. We therefore examined the binding specificity of all probes giving significant eQTL results (discussed in section 4.5 Quality control of the data). Secondly, microarray only allows the assessment of known transcripts [140]. The Illumina HumanWG-6 v3 microarray was designed some time before the 21^{rst} of May 2008, as this is the date when the microarray probe information was published in the Gene expression Omnibus. Therefore we have not obtained RNA level measurements from transcripts identified after this date . Lastly, microarrays cannot distinguish between transcript isoforms, unless they have been specially designed. Distinction between mRNA isoforms would have been valuable to us in order to get an indication about whether particular transcripts were causing the difference in gene expression observed in the eQTLs.

4.4.2 RNA Sequencing

While microarray allows us to assess the RNA content and abundance for all known transcripts, RNA sequencing extends the scope and depth of investigation to the entire transcriptome of known and novel transcripts [141]. In **paper III**, we used high throughput RNA sequencing to assess the transcriptome of the individual thymic APCs.

In practice, RNA is converted to a library of cDNA fragments with adaptors attached to one or both ends [140]. When RNA yields are low, each cDNA molecule can be amplified by PCR before adaptors are attached. Then the cDNA molecules are sequenced in a high-throughput manner to obtain short sequences from one end (single end sequencing) or both ends (paired end sequencing). The length of the these short sequences called “reads” varies between 30 – 400 bp, depending on the sequencing technology used [140]. Reads can then be computationally mapped to a reference genome to reveal a transcriptional map, where the number of reads aligned to each gene gives a measure of its level of expression [142]. Common computational analysis steps in RNA sequencing experiments include quality control of the reads, read alignment, assigning reads to genes or transcripts, and estimating gene or transcript abundance [141]. The RNA sequencing procedure used for the APCs is described in detail in **paper III**, and only certain aspects about the method will be discussed below.

4.4.2.1 Preparation for RNA sequencing

Laboratory protocols for the RNA extraction method, cDNA amplification, indexing and library sequencing protocol are critical for obtaining good sequence data. As mentioned earlier, the APC samples used in **paper III** suffered from small cell numbers (Table 2) with low RNA yields (Table 3). Luckily, a variety of amplification-based methodologies have been proposed to handle the issues with low RNA input [143]. We chose to use the SMART-Seq v4 Ultra Low Input RNA Kit for Sequencing (Clontech Laboratories), as this kit accepts RNA yields down to 10 pg. SMARTer-sequencing technology is based on associating universal primer sequences to either ends of the cDNA library followed by global PCR amplification of all transcripts by using complementary sequences of the universal primers [144]. We used 100 pg of RNA from each APC sample to generate and amplify cDNA.

SMARTer technology exhibit high transcriptome coverage across different amounts of mRNA input (1 ng, 100 pg, 50 pg, 25 pg) when it is compared to “standard” RNA sequencing (Std. RNA-Seq), using 50 ng of mRNA and no amplification [143]. Approximately 80% of the transcriptome was shown to be covered when using 100 pg and the amplification-based SMARTer kit [143]. This suggests that we have covered nearly 80% of the in vivo transcriptome of the APCs in **paper III**. Furthermore, the distribution of mapped reads across the length of the transcript was also reported to overlap for SMARTer technology and Std. RNA-Seq [143]. However, one limitation of SMARTer technology is that it does not efficiently amplify transcripts longer than 4 kb [145]. Long transcripts in libraries treated with

SMARTer technology exhibited lower read counts in comparison to Std. RNA-Seq [143]. This indicates that transcripts over 4 kb in our APC libraries might suffer from length bias. Among the genes investigated in **paper III**, we found 16 genes with transcripts in the Ensembl database (GRCh38) with protein-coding transcripts over 4 kb: *NRP1*, *XCR1*, *THBD*, *ITGAX*, *CANX*, *ERAP1*, *ERAP2*, *TAP2*, *CTSS*, *FCRL3*, *AHI1*, *ELMO1*, *GPR65*, *IP6K1*, *SLC16A14* and *TROVE2*. However, among these genes, 9 were < 5000 bp, and only *TROVE2* had a transcript > 6000 bp (n = 7848 nt).

The final cDNA yields we obtained from the SMART-Seq kit are listed in Table 4.

Table 4: Median cDNA yields and IQR of the thymic APCs

APC	**Median cDNA yield (IQR)**
CD123+ DC	14.4 ng (8.17)
CD141+ DC	10.6 ng (8.40)
CD19+ B cells	18.2 ng (5.68)
mTEC	8.2 ng (6.69)

Finally, 1 ng from each APC sample was further indexed with MicroPlex Library Preparation Kit v2 (Diagenode) before the samples were delivered to the Norwegian Sequencing center (NSC). Longer reads increase the level of uniquely mapped reads and paired end sequencing has a higher alignment rate. We therefore chose 125 bp paired end sequencing of the APC samples, which was performed on an Illumina HiSeq 2500 (Illumina, CA, US) instrument with 4 samples per lane.

After sequencing, we preprocessed and mapped the reads to a genome and a transcriptome reference (GRCh38), before quantifying the data. Because reads can be unambiguously mapped to unique regions of the genome, RNA sequencing has very low, if any, background noise signal compared to microarray [140]. We used STAR [146] to align and map the reads from the APC samples. STAR defines the percentage of uniquely mapped read, or mapping rate, as the proportion of uniquely mapped reads out of all input reads [147]. A library is defined as "very good library" if the uniquely mapped read percentage exceeds 90%, and as "a good library" if it exceeds 80% [147]. In **paper III**, 21 out of 23 APC samples had uniquely mapped read percentages > 90%. One CD19+ B cell and one mTEC sample contained 83% and 72% uniquely mapped reads, respectively. However, mapping rates need to be below 50% to indicate a problem with the library preparations or the data processing

[147]. The two samples with the lowest uniquely mapped read percentage had among the highest percentage of duplicates.

As duplicate removal is a common practice in next generation sequencing pipelines [148], duplicate aligned reads in our libraries were removed with Picard Tools, MarkDuplicates. Artificial duplicates can arise from amplification during sample preparation [148]. If Smart-sequencing technology is used, these PCR duplicates are identified computationally as read duplicates [149]. However, such read duplicates can also arise by sampling independent molecules. The chance that the latter type of read duplicates, or "natural duplicates", occur for a transcript of a given length, increases with expression levels [149]. The danger about removing duplicates bioinformatically is therefore that the program can incorrectly mark identical reads arising from a highly expressed gene as PCR duplicates arising from the library preparation. The MarkDuplicates tool reported both paired and unpaired read duplicates in our libraries (Table 5). These duplicates might be PCR duplicates that have arisen during amplification with the SMARTer kit or the MicroPlex preparation kit; however, they could also be expression reads. Most likely, it is a combination. However, in the latter case, we might have lost some valuable information from highly expressed genes in the APCs in **paper III**.

Table 5: Median number of duplicates from paired and unpaired reads and IQR of the thymic APCs

APC	**Median number of read pair duplicates (IQR)**	**Median number of unpaired read duplicates (IQR)**
CD123+ DC	21,386,435 (5,660,228)	192,805 (52,040)
CD141+ DC	23,469,675 (4,179,364)	173,367 (39,834)
CD19+ B cells	24,999,741.5 (4,588,379)	240,029.5 (29,750)
mTEC	30,762,761 (8,435,145)	182,245 (78,482)

Finally, paired reads were counted with featureCounts. The featureCounts files were then further processed with edgeR. edgeR is a Bioconductor software package for examining differential expression of replicated count data [150]. In short, the steps in edgeR involved filtering lowly counted reads (less than 1 count-per-million (CPM)) and genes expressed in fewer biological replicates than the number of biological replicates in to smallest group. Then the library was normalized, and a multidimensional scaling plot (MDS) plot was made in order to examine the samples for outliers. The MDS plot is further discussed in Section 4.5

(Quality control of the data). Furthermore, as the group with the smallest number of biological replicates in **paper III** was the mTECs (n = 5), we used a threshold deciding that genes need to be present in 5 biological replicates to be included in the APC dataset. However, this threshold, which is intended for avoiding false positive findings, actually led to a bias in one of our analyses (**paper III**). When we addressed the number of TRA genes expressed in the individual APC, this threshold would in fact restrict the number of TRA genes, because TRAs are only expressed in a minority of mTECs (1-3%) at any given time [72]. We therefore generated a separate dataset where we lowered the threshold to 1. However, this further leads to a new problem, because in addition to lowly expressed TRA genes, we also increase the risk of including false positives in the APCs. Nevertheless, as discussed in the article, it is not possible to know whether the lowly expressed transcripts are in fact processed to proteins and further to peptides presented on the cell surface. More studies are needed to confirm the peptide repertoire in the HLA molecules on the thymic APCs, which is currently difficult to achieve due to the limited available technology today.

4.5 Quality Control of the data

In **paper I and paper II**, because of the vast number of SNPs interrogated on the Imunochip (n = 195806) and the number of probes on the microarray (n = 48802), we performed quality control of SNPs and probes after testing the AID associated risk SNPs for eQTL associations with surrounding genes (+/- 1 Mb).

In the list of SNP-probe tests, we controlled that the AID risk SNPs (seven in **paper I** and 393 in **paper II**) had a genotype success score (GSS) > 90%. In **paper I**, one SNP (rs30187) obtain a GSS of 81%, this SNP was therefore retyped by Taqman allele discrimination assay (Thermo Fisher Scientific, USA) to obtain a 100% GSS. In **paper II**, we found two SNPs (rs941576 and rs12261843) that had a GSS = 54.8% and GSS = 0%. We also found one variant (rs1156425) that had 100% GSS but was not polymorphic in our thymic population. The SNP-probe pairs including these three risk SNPs were excluded from the study. In order to test for population stratification, we also verified that all SNPs had Hardy-Weinberg equilibrium (HWE) P-values > 0.05. In **paper II**, we found nine SNPs that had HWE P-values < 0.05, these were therefore excluded. Furthermore in **paper II**, two SNPs were located on the on the X chromosome and was excluded to avoid gender as a confounding factor. SNP-probe pairs in the HLA region (Chromosome 6: 29,705,659-33,817,929 in GRCh37) were also excluded, as this gene region is highly polymorphic. Only a few HLA probes are present on the microarray and the whole repertoire of alleles at each gene is therefore not represented. Studies using microarrays have reported differences in the gene expression levels of HLA alleles between individuals [151], however, we suspect that the high number of polymorphisms causes technical problems and makes it difficult to obtain reliable data. Therefore, 26 SNPs in the HLA region and the corresponding SNP-probe pairs were therefore excluded. In the end, seven (**paper I**) and 353 (**paper II**) AID risk SNPs were tested against surrounding probes (+/- 1 Mb), resulting in 215 (**paper I**) and 15843 (**paper II**) SNP-probe tests.

We also performed quality control of the probes. In **paper I** and **II**, because of the risk of cross-hybridization on the microarray, we examined the binding specificity of all probes giving significant eQTL results. In **paper I**, the eProbe ILMN_1752145 was binding completely (50 bp) to exon 20 in of *ERAP1*, but also to 24 bp of exon 21 in *CAST*. The binding of 24 bp is most likely not stable enough to resist the washing step after the microarray hybridization. However, we cannot completely exclude the possibility that we have obtained signals from *CAST* transcripts on this probe. Furthermore, in **paper II**, we found one unspecific eProbe (ILMN_2209027) which bound to both *RPS26* and *COL4A3BP*, and was therefore excluded from the study. Finally, as probes covering common SNPs must be interpreted with caution in eQTL analyses [152], the probe sequences underlying the significant eQTLs in both **paper I** and **II** were aligned against the University of California Santa Cruz (UCSC) GRCh37 and investigated for overlap with common SNPs (MAF) > 1%). In **paper I**, one SNP, rs27044, was found within the probe sequence of ILMN_2336220 binding to *ERAP1*. This SNP was also in high LD with two other risk SNPs that we tested, rs30187 (r^2 = 0.7) and rs27432 (r^2 = 0.4). We assessed whether these eQTLs were true by measuring the *ERAP1* expression with the Taqman gene expression assay Hs_00429970 (Thermo Fisher Scientific. The eQTL associations were found to be non-significant for both rs30187 (P-value > 0.24) and rs27432 (P > 0.11), hence the SNP-probe pairs comprising these SNPs were excluded.

We should have removed the probes binding non-uniquely or overlapping with common SNPs (MAF > 1%) before the eQTL testing. However, because of the number of probes on the microarray (n = 48802), this seemed like a complicated task at the time. Today I know that a programmed script could have checked which probes that mapped uniquely to exons in the human reference, and whether they interfered with common SNPs (MAF > 1%). This could have eliminated spurious SNP-probe pairs before we tested for eQTL associations.

In **paper III**, a MDS plot was made as a quality control in order to examine the APC samples for outliers. A MDS plot of the samples represent the distances correspond to leading log-fold-changes between each pair of samples. The leading log-fold-change is the average (root-mean-square) of the largest absolute leading log-fold-changes between each pair of samples. The MDS plot revealed that one of the mTEC samples has been contaminated by DCs, as it was the only "mTEC" sample physically located among the CD141 DC samples (Figure 10). This sample was therefore excluded (**paper III**).

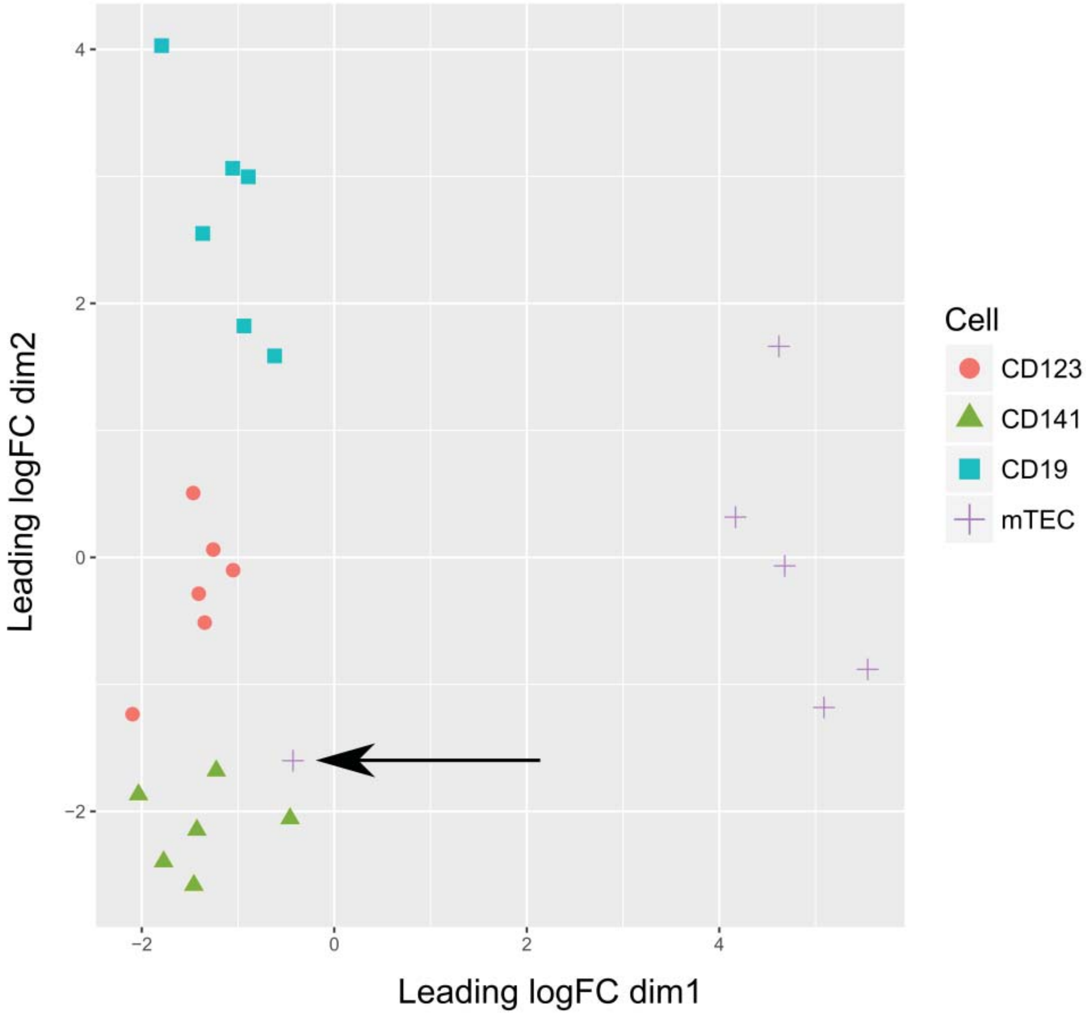

Figure 10: MDS plot displaying the leading log-fold-change distance between the thymic APC samples. One mTEC sample was contaminated by DCs in the lab, so this sample clearly changed cluster (marked by arrow).

4.6 Statistics

In the three papers of this thesis, we have performed comparisons between groups to detect significant differences in gene expression levels. In this regard, hypothesis testing is an essential procedure in statistical validation. The null hypothesis, H_0, states that there is no true difference between the two groups compared, and is either confirmed or rejected based on the P-value and the significance level of the experiment. The P-value is the probability of obtaining a result at least as extreme as the one observed in the sample data, assuming H_0 is true. If for example a study reports a difference with a P-value of 0.04, then this P-value indicates that if there were truly no difference, you would still obtain the observed difference (or a larger one) in 4% of the same studies due to random sampling error. H_0 is rejected if the P-value is less than a predefined significance level α. The significance level, denoted as alpha or α, is therefore the probability of rejecting the null hypothesis when it is true. If for instance the significance level is set at $\alpha = 0.05$, then there is a 5% chance of finding a difference which in fact is a false positive (a type I error) and concluding that this difference exists when there is no actual difference.

Depending on your data, different types of statistical tests can be used. In the eQTL analyses (**paper I and paper II**), the aim was to test whether there was a significant difference in gene expression of a given gene between the genotype groups of a specific SNP (A/A, A/B, B/B) in the 42 individuals. P-values were given by the Wald test, which is parametric statistical test used in PLINK for quantitative trait associations [153]. In **paper III**, we tested for differential gene expression and differential exon usage between the APCs. These tests were performed in edgeR, which uses a generalized linear models (GLM) and GLM likelihood ratio tests to determine differentially expressed genes between the cell types trough pairwise comparisons [150]. Differential exon usage (called differential splicing in **paper III**) analysis was performed by applying the edgeR F-test.

The significance level α is dependent on the number of tests (n) performed. A higher number of tests will require a stricter or more conservative significance level in order to increase the probability of detecting a true difference. In **paper I and II**, to adjust for multiple comparisons, we applied conservative Bonferroni correction, which can be set to $\alpha = 0.05/n$. It can be discussed whether Bonferroni is too conservative, because this type of correction assumes complete independence between tests. This assumption is not reasonable when LD exists between the markers tested. This increases the possibility for type II errors, i.e. falsely accepting the null hypothesis. For instance, in **paper I**, we detected two eQTLs where the eSNPs (13003464 and rs10181042) were both associated with the same eGene (C2orf74). When we performed conditional analysis, the eQTL signal from rs10181042 displayed the strongest statistical significance, whereas rs13003464 obtained a P-value of 0.39, indicating that rs13003464 was simply a dependent signal. In this case, these two SNPs should not be counted as two individual tests, which further have contributed to the (too) strict threshold in our study. Among the 353 GWAS SNPs tested in **paper II**, there may also be several more SNPs in LD, however this was not taken into account when calculating the significance level α. We might therefore have lost true eQTLs in our screens (**paper I and II**). In **paper III**, the P-values were adjusted for false discovery rate (FDR). FDR is designed to control the expected proportion of false positives. This procedure is a less stringent control of type I errors compared to Bonferroni correction, which controls the probability of at least one type I error. A significance cut off at 0.05 was used for the FDR adjusted P-values.

The power, or sensitivity, of a test is the probability that the test correctly rejects the null hypothesis when it is false. Power is dependent on the sample size, the magnitude of the effect of interest in the population and the statistical significance level used in the test. In **paper I and paper II**, we had a limited sample size of thymic tissues (n = 42), and therefore, presumably, we did not have the power to detect all AID-eQTLs. If more individuals were included in these studies, more genotypes and gene expression levels would have been added to the SNP-probe tests, resulting in more robust P-values, possibly helping more SNP-probe pairs to pass the strict significance thresholds (3.1 x 10^{-6} and 2.7 x 10^{-5}) that was set to avoid false positive results.

Likewise, given the limited sample size we did not have the power to perform conditional analyses in order to fine map the location of the lead eSNPs causing the eQTL signals (the eQTL "peaks"). If we had a larger sample size, the robustness of our observation would also have increased, and there might potentially be other eSNPs that would obtain more significant P-values than the novel lead SNPs we found (*ERAP1* – rs7063, *ERAP2* – rs27302, *RNASET2* – rs429083, *SYS1* – rs2743414, *AJ006998.2* - rs991774). For instance, in **paper I**, we noticed that rs27302 seemed to be the sole SNP that correlated with the highest *ERAP2* expression level (Figure 11).

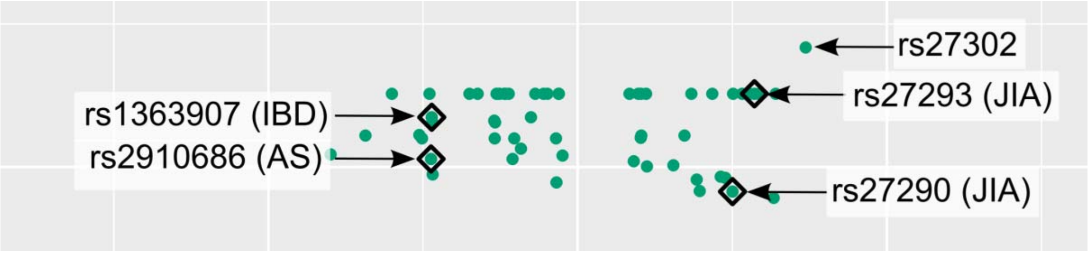

Figure 11: A section cut from Figure 1 in **paper I** showing the LD block with SNPs that obtained the most significant P-values with the *ERAP2* probe ILMN_1743145.

However, rs27302 belonged to a large LD block ($r^2 > 0.9$), including the AID risk variants rs2910686, rs1363907, rs27290 and rs27293. The superior P-value of rs27302 turned out to be caused by the genotype of one individual in our sample, reflecting the influence of small numbers. When we investigated the genotypes for the AID risk variants and the lead SNP (in the order rs2910686 - rs1363907 - rs27290 - rs27293 - rs27302) that correlated with the highest *ERAP2* levels in our population, we found five individuals with the haplotype G-A-G-A-G, and one individual with the haplotype G-A-G-A-A. In order to test whether it is in fact rs27302 that is the real lead eSNP, we could theoretically have performed conditional analysis, conditioning on rs27302, to see whether any of the other AID SNPs would obtain a non-significant P-value (indicating that it is in fact rs27302 that is the lead eSNP) or if they remained statistically significant (indicating that another SNP is the lead eSNP). However, to do so we would have needed more than one individual with the deviating G-A-G-A-A haplotype in order to understand whether it is in fact rs27302, and not any of the other SNPs, that represents the causal eSNP.

4.7 The Human Protein Atlas

In **paper III**, we used the Human Protein Atlas to search for TRA genes. Mapping the human proteome in all the organs of the body and defining proteins that are “tissue-specific” has been a major goal in the research community as this will greatly increase our knowledge of human biology and disease [154].

The Human Protein Atlas portal (www.proteinatlas.org) provides a map of the human tissue proteome based on quantitative transcriptomics on tissue and organ level combined with protein profiling using microarray-based immunohistochemistry to achieve spatial localization of proteins down to the single cell level [154]. The authors behind this project have analyzed all major tissues and organs (n = 44, thymus was not included) in the human body using 24028 antibodies (producing 13 million immunohistochemistry images) as well as RNA sequencing data for 32 of the 44 tissue types. Based on this, the authors classified 20334 putative protein-coding genes into categories based on their expression levels across the 32 tissues. 34% of the protein coding genes showed elevated expression levels in at least one of the analyzed tissues [154], and were therefore categorized as either:

(i) Tissue enriched, if the gene was expressed (> 1 transcript per million (TPM)) in one tissue at least five-fold higher than all other tissues

(ii) Group enriched, if the gene had a five-fold higher average TPM in a group of two to seven tissues compared to all other tissues

(iii) Tissue enhanced, if the gene had a five-fold higher average TPM in one or more tissues compared to the mean TPM of all tissues.

Furthermore, if a gene had more than 1 TPM in all tissues, the authors termed this gene as "expressed in all". If a gene had less than 1 TPM in all tissues, the gene was "not detected". Lastly, if a gene was detected in at least one tissue but was not belonging to any of the categories above, the gene was "mixed". A pie chart of the number of genes in each category is shown in Figure 12.

The largest number of "tissue-enriched" genes was found in the testis, subsequently followed by the brain and the liver [154]. An important point in this study is that the use of the word "tissue- specific" has been avoided because this definition depends on arbitrary cut-off levels. Many proteins described in the literature as "tissue-specific" were in fact shown to be expressed in several tissues [154].

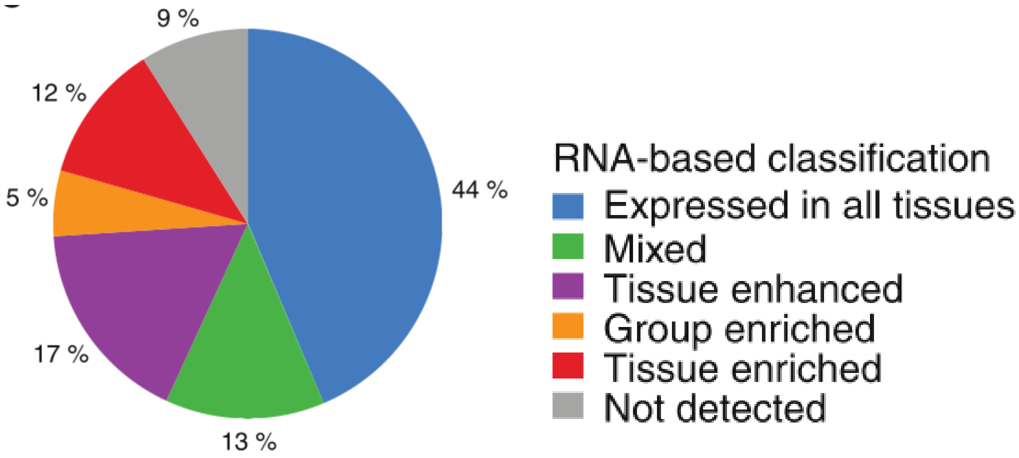

Figure 12: Number of genes classified in each expression category. Figure from [154] with permission.

In **paper III**, we downloaded the list 2608 "Tissue enriched" genes from the Human protein Atlas, and defined these as "TRA" genes in our study. We further searched for these genes among the thymic APCs. We found expression (FPKM > 1) in all four APC types, even in the CD141+ DCs and the CD123+ DCs, although the two latter has never been reported to transcribe TRA genes before. However, as the tissue-enriched genes are not strictly tissue specific, it is maybe not so surprising that we detected TRA gene expression levels in these cell types. To address whether these selected TRAs are actually presented by the APCs in thymus in order to negatively select thymocytes reacting towards them, a link from the RNA levels seen in the APC and a the peptide repertoire presented by the thymic APCs needs to be generated. As mentioned in **paper III**, these types of experiments are currently limited by available technology [155, 156].

5. Discussion

This thesis addresses AID risk variants that influence the gene expression profile in whole thymus and the bioinformatically assessed functional role of these variants. Furthermore, we have investigated the transcriptomes of four thymic APCs, with emphasis on genes important for APC function and genes of relevance to AID. In this chapter, we have explored whether any of the thymic AID-eQTLs can be mapped back to the four thymic APCs.

5.1 Are the thymic APCs affected by the AID-eQTLs?

The findings in this project suggest that risk variants associated with AIDs affect gene regulation in the thymus organ. The question is which thymic cell types are affected, and which cell mechanisms are influenced by the variation in gene expression. Although the thymocytes make up approximately 98% of the cells in human thymus [102], the thymic APCs have a pivotal role in the fate choices of developing T cells [81] to prevent autoimmunity, and therefore we investigated the expression levels of the eight eGenes discovered in **paper I** and **paper II** in the thymic APCs (**paper III**). An overview of the thymic eGenes and findings from the three papers are summarized in Table 6, Figure 13 and 15.

Table 6: The thymic eGenes and findings from the three papers.

eGene	Gene description	Thymus-specific eQTL*?	RNA tissue specificity **
AJ006998.2	LincRNA	Yes	-
C2orf74	Chromosome 2 open reading frame 74	No	Expressed in all
ERAP2	Endoplasmic reticulum aminopeptidase 2	No	Mixed
FCRL3	Fc receptor like 3	No	Grouped enriched
NPIPB8	Nuclear pore complex interacting protein family member B8	Yes	Tissue enriched
RNASET2	Ribonuclease T2	No	Expressed in all
SIRPG	Signal regulatory protein gamma	No	Tissue enhanced
SYS1	Golgi trafficking protein	No	Expressed in all

*The publically available eQTL resources employed to investigate whether our significant SNP-probe pairs are thymus-specific include the GTEx Portal, GTEx eQTL Browser and Haploreg V4 (supplementary table S8 in paper II). ** RNA tissue specificity was addressed using the Human Protein Atlas.

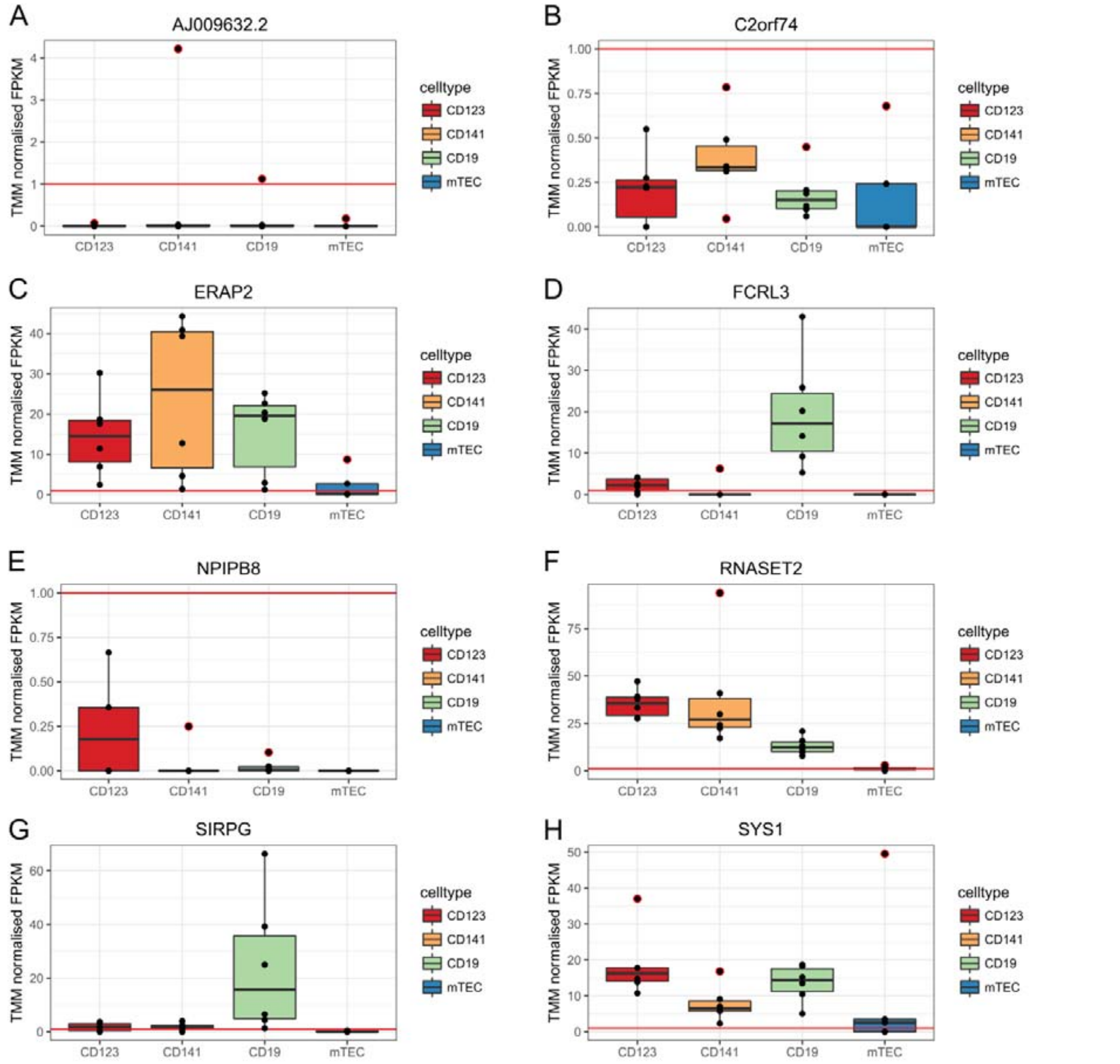

Figure 13: Expression levels of the eGenes (**paper I** and **II**) in the individual APCs (extracted from **paper III**). Boxplots represent the median and quartiles of the relative RNA expression levels as normalized FPKM. The X-axis shows the individual thymic APCs and the Y-axis shows the TMM normalized FPKM. In paper III, “expressed genes” were defined as genes with an FPKM >1 (red line). Black dots represent the individual biological replicates. Black dots encircled in red are outliers. A) AJ009632.2 (or ENSG00000229425.2, the non-coding DNA sequence from where the transcript AJ006998.2, or ENST00000634644.1 derives) B) C2orf74; C) ERAP2; D) FCRL3; E) NPIPB8; F) RNASET2; G) SIRPG; H) SYS1. The dataset used here includes genes with a CPM > 1 and presence in at least one biological replicate.

Five of the eight eGenes (*ERAP2*, *FCRL3*, *RNASET2*, *SIRPG* and *SYS1*) were expressed with an FPKM > 1 in at least one APC. *AJ009632.2*, *C2orf74* and *NPIPB8* were not expressed according to our expression threshold (FPKM = 1), albeit low levels (FPKM < 1) could be observed in some of the thymic APCs. To further understand the function of the eGenes, and the potential effect of the eQTLs on cell mechanism, I will address each eGene individually, and integrate the findings from the three papers.

5.2 The individual eGenes

5.1.1 ERAP2

The eGene *ERAP2* was implicated in the most significant eQTL ($P < 8.22 \times 10^{-23}$) that emerged from our eQTL screens. This gene encodes endoplasmic reticulum aminopeptidase 2 belonging to the M1 family of zinc-metallopeptidase enzymes [157]. This enzyme trims peptides to optimal length for binding in the MHC class I molecules before presentation on the cell surface to CD8+ T cells. *ERAP2* gene function has not been extensively studied because of its absence in rodents. However, in 2010, Andrés et al. studied blood samples from six human populations and reported that *ERAP2* has evolved under balancing selection, maintaining two major haplotypes (haplotype B and A) with frequencies 0.56 and 0.44, respectively [158]. Similarly to Andrés et al., we also observed two major haplotypes in the genotype dataset from the 42 thymic tissues (**paper I**). The haplotype with the largest frequency (0.619 in our study population) included the A allele of the thymic eSNP (rs27302) and the protective alleles of the AID SNPs (rs2910686-A, rs1363907-G, rs27290-A and rs27293-G) in high LD ($r^2 > 0.94$). Conversely, the second most frequent haplotype (0.369) comprised the rs27302-G allele and the risk alleles of the AID SNPs (rs2910686-G, rs1363907-A, rs27290-G and rs27293-A).

Andrés et al. further reported that the *ERAP2* transcript expressed by haplotype B (individuals carrying the rs2248374 G-allele) undergoes differential splicing, resulting in an isoform variant with an extended exon 10 and a premature stop codon. This mRNA is assumed to undergo nonsense-mediated decay, as (almost) no protein could be detected by western blot in BB homozygotes [158]. In **paper I**, we noticed that there was moderate to high LD between the splice SNP rs2248374 from Andrés et al. and our thymic eSNP rs27302 ($r^2 > 0.66$ in our genotype data and $r^2 > 0.82$ in the dataset from the 1000 Genomes phase 3). We also observed that the eSNP rs27302 resulted in a more significant eQTL ($P = 8.22 \times 10^{-23}$) than rs2248374 ($P = 2.74 \times 10^{-9}$), and when we conditioned on rs27302, rs2248374 obtained a non-significant P-value (P = 0.22). This suggest that the *ERAP2* expression in thymus is first and foremost regulated by rs27302 (or another SNP that has not been tested), and secondly by rs2248374. However, this needs to be confirmed by further analysis.

Finally, Andrés et al. further performed flow cytometry analysis and showed that the standardized mean fluorescence intensities from HLA-class I (HLA-A, -B and -C) molecules was significantly lower on the CD19+ B cell surface in BB homozygotes (who carry the rs2248374-G allele) [158]. In **paper I**, we found that the lowest gene expression detected for *ERAP2* was associated with the rs27302-A allele. If individuals homozygous for the rs27302-A allele and the rs2248374-G allele (the two alleles with the highest frequencies in the population) express the aberrant *ERAP2* mRNA, this could potentially indicate that these individuals also have lower numbers of HLA-class I molecules on the cell surface. Conversely, individuals homozygous for the rs27302-G allele (often located on the same haplotype as all the AID SNP risk alleles) and the rs2248374-A allele would express the mRNA encoding a full-length ERAP2 protein.

We questioned whether the relationship between *ERAP2* expression and HLA-class I gene expression could be observed in our APC data (Figure 14). Although there was no clear relationship between the expression levels of *ERAP2* and *HLA-A*, *HLA-B* or *HLA-C*, we did notice that the two individuals (82 and 83) with the lowest levels of *ERAP2* also consistently had among the lowest levels of the HLA class I genes. However, more than six individuals are needed to confirm this hypothesis. Furthermore, edgeR does not recommend comparing the expression levels from two genes in one cell type, as edgeR is only concerned with relative changes in expression levels between conditions, and not directly with estimating absolute expression levels [159].

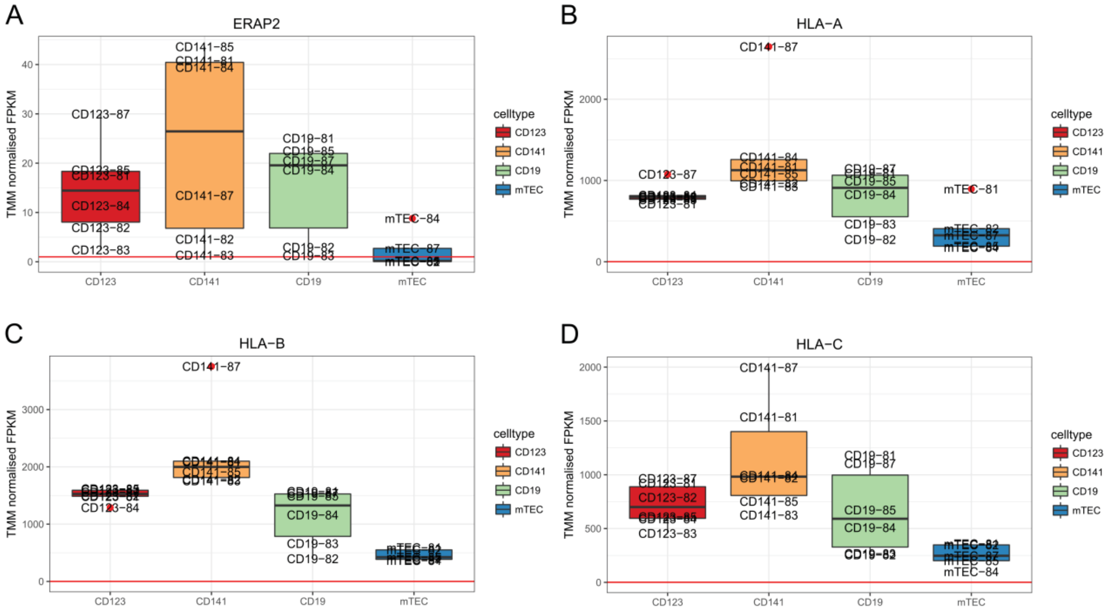

Figure 14: Analysis of the relationship between *ERAP2* and the HLA-class I genes A. *ERAP2* B. *HLA-A*, C. *HLA-B* and D.*HLA-C*. Boxplots represent the median and quartiles of the relative RNA expression levels as normalized FPKM. The X-axis shows the individual thymic APCs and the Y-axis shows the TMM normalized FPKM. Black dots are replaced by the IDs of the individual biological replicates. Black dots encircled in red are outliers.

Taken together, it seems to me that little or no *ERAP2* expression is beneficial, whereas having *ERAP2* expression is associated with AID risk, possibly because of more HLA class I molecules on the cell surface. One could further speculate whether it is in fact sufficient to have ERAP1? Not many studies have addressed the relationship between the amounts of MHC class I molecules and predisposition to autoimmunity, but an interesting hypothesis, which has been supported by others, states that there is a link between HLA class I expression and the pathogenesis of AID [160, 161]. HLA class I is expressed on all nucleated cells, meaning that all cell types in the human thymus might be affected by the variation in *ERAP2* expression. This could possibly have an implication on thymocyte development or self-antigen presentation by thymic APCs. Further studies on how variation in *ERAP2* expression influences these cell types will be needed in the future to answer these questions.

5.1.2 ERAP1

We found no clear link between the most significant eSNP (rs7063) and the AID risk variants (rs27432 and rs10045403) in *ERAP1* because of the low LD ($r^2 = 0.22$ and $r^2 = 0.37$, respectively). The eQTL signals from the AID SNPs were also markedly inferior. It is therefore likely that the AID risk variants at *ERAP1* exert their effect in other ways than influencing thymic *ERAP1* gene regulation. These findings show the importance of thoroughly mapping eQTL signals in tissues, and that although a risk variant correlates with an eGene, it doesn't necessarily mean that the risk variant is the causal eSNP. However, we cannot exclude the possibility that other thymic eSNPs in higher LD with rs27432 and rs10045403 would be detected if a larger number of thymic tissue samples were analysed and more power was added to the study.

5.1.3 SIRPG

Which allele of the eSNP rs2281808 that is associated with T1D risk is not reported in the GWAS catalog [136]. It was therefore not possible to see which way the risk allele influenced the expression of *SIRPG*. *SIRPG* encodes the signal-regulatory protein gamma (CD172G), a member of the signal regulatory protein family [162]. Unlike its family members SIRPα and SIRPβ, SIRPγ is highly expressed in SP CD4+ and CD8+ T cells [162]. SIRPγ on the T-cell surface binds to CD47 on other cells to increase cell-cell adhesion [162, 163]. Furthermore, interaction between SIRPγ on T cells and CD47 on APCs promotes antigen-specific T-cell proliferation and costimulates T-cell activation [163]. Interaction with APCs in thymus is an extremely important process for the developing T cells, as this is the way their TCR is tested for functionality and self-peptide recognition. It is therefore conceivable that the thymic *SIRPG* eGene from **paper II** is expressed in the developing thymocytes. Moreover, SIRPγ is also expressed in B cells [164]. Consistent with this, we detected *SIRPG* expression in CD19+ B cells among the thymic APCs. Low levels were also detected in both DC subsets. The eSNP underlying the *SIRPG* eQTL was originally an intronic SNP from GWAS associated with type 1 diabetes, but then the disease signal was further fine-mapped to a missense variant (p.Val263Ala) in exon 4 of the gene [42]. Valine and alanine both have aliphatic R groups that are non-polar and hydrophobic, it is therefore difficult to interpret the functional impact of this amino acid change on SIRPγ, and how it eventually contribute to AID susceptibility.

5.1.4 FCRL3

Homozygosity for the eSNP rs3761959-G risk allele in the *FCRL3* eQTL led to a decrease in gene expression. This eSNP was associated with MS and GD, and in 2013, RA was also found associated with *FCRL3* [45]. *FCRL3* encodes an Fc receptor-like glycoprotein. Fc receptors bind to the Fc portion of immunoglobulins and can trigger phagocytic or cytotoxic cells to destroy microbes. The Fc receptor-like genes have similar features to Fc receptor genes [165]. FCRL3 has both a immunoreceptor tyrosine-based activation motifs and a immunoreceptor tyrosine-based inhibition motifs and has therefore been suggested to have an activating/inhibitory or a fine-tuning role in regulation of immunologic function [165]. When

FCRL3 expression was investigated in cell lines representing different hematopoietic lineages, it was only found in mature B cell lines [165]. Consistent with this, expression from the *FCRL3* eGene was clearly observed in the thymic CD19+ B cells, and low levels were also detected in the CD123+ DCs. It is therefore imaginable that the eQTL influence the fine-tuning role of FCRL3 in thymic CD19+ B cells, and possibly also in CD123+ DCs, perhaps in relation to central tolerance.

5.1.5 RNASET2

We observed a higher gene expression of *RNASET2* in the thymus of individuals that were carriers of the eSNP rs415890-C risk allele, associated with CD. GWAS has in addition reported *RNASET2* as a risk factor for vitiligo [166]. *RNASET2* encodes the ribonuclease T2, and is a human member of the Rh/T2/S family of acidic hydrolases [167]. Ribonucleases are ubiquitous, conserved enzymes that are involved in RNA metabolism [167]. Interestingly, the involvement of RNASET2 in the pathogenesis of vitiligo has been addressed [168]. Vitiligo is an AID where melanocytes are destroyed, resulting in depigmented skin [166]. A study by Wang et al. [168] focused on the fact that RNASET2 is secreted from the cells under stress conditions (by for instance ultraviolet irradiation, mechanical injuries and inflammation). This was supported by their finding that RNASET2 was overexpressed in the epidermis of vitiligo patients compared to healthy controls [168]. Their hypothesis further is that RNASET2 might act as an endogenous ligand that activates APCs and further leads to an immune response against melanocytes [168]. Another study has shown that omega-1, an RNaseT2 family member secreted from the eggs of *Schistosoma mansoni*, could induce Th2 polarization through DC priming [169], and therefore Wang et al. will investigate further whether or not the human RNASET2 has similar effects [168]. According to the Human Protein Atlas, expression of *RNASET2* can be detected in all tissues, and among the thymic APCs, we could also detect *RNASET2* expression in all APCs, with the highest levels in CD123+ and CD141+ DCs. It is of course imaginable that *RNASET2* is expressed in thymic tissue as a response to stress (for instance because of the cardiac surgery). However, it could also be expressed ubiquitously due to its role in RNA metabolism. Nevertheless, there seems to be a link between a higher expression of *RNASET2*/RNASET2 and AID. More studies are needed to understand the impact of the thymic *RNASET2* eQTL in DCs, and possibly in thymocytes.

5.1.6 SYS1

The Ps associated eSNP rs1008953-C risk allele correlated with a higher gene expression of *SYS1*. Although few functional studies have been performed with the SYS1 protein, one study reports that the homologue of SYS1 in Saccharomyces cerevisiae, sys1p, is a Golgi membrane protein that possibly forms a complex with the Arf-like GTPase 3p (Arl3p) and targets Arl3p to the Golgi apparatus [170]. The human homologue of Arl3p, ADP-ribosylation factor-related protein 1 (ARFRP1) co-precipitates with the human homologue of sys1p by chemical cross-linking [170], suggesting a similar role in humans. A model that emerged from these findings suggest that sys1p recruitment of Arl3p will further, through progressive steps involving other factors, lead to the capture of vesicles from the endosomes and fusion with the trans-Golgi network [171]. Consistent with the fact that *SYS1* is expressed in all tissues according to the Human Protein Atlas, *SYS1* was also expressed in all four thymic APCs. The slightly higher *SYS1* eGene expression in individuals homozygous for the rs1008953 risk allele could possibly affect the trafficking of vesicles from endosomes to the Golgi. In the endosomes, processed antigenic peptides and HLA class II proteins are present [172], it is therefore tempting to speculate that the variation in *SYS1* expression could perhaps indirectly interfere with the HLA-class II pathway in the APCs. However, further studies are needed to understand the functional consequence of the thymic *SYS1* eQTL.

5.1.7 NPIPB8

The *NPIPB8* eQTL comprised the eSNP rs151181, where the C risk allele was associated with CD and a higher *NPIPB8* expression. This eQTL was one of the two potentially thymus-specific eQTLs in **paper II**. *NPIPB8* encodes member B8 of the nuclear pore complex interacting protein family. According to the EMBL-EBI database (http://pfam.xfam.org/family/npip), the function of this family is unknown. In the Human Protein Atlas, *NPIPB8* is a tissue-enriched gene expressed in the testis, but it is also detected at low levels in the small intestine, duodenum, ovaries, adrenal gland, endometrium, epididymis, stomach, cerebral cortex, appendix, prostate, spleen, skin and 15 other tissues (https://www.proteinatlas.org/ENSG00000255524-NPIPB8/tissue), suggesting that NPIPB8 is an ubiquitous, but lowly expressed protein. Among the thymic APCs, *NPIPB8* was not expressed (FPKM > 1), albeit low levels could be detected in the CD123+ DCs. This suggests that *NPIPB8* might be an eQTL in thymocytes. More studies are needed to understand the function of this family and particularly this member, and to further investigate the functional implication of the potential thymus-specific *NPIPB8* eQTL.

5.1.8 C2orf74

In *C2orf74*, we found a higher gene expression in individuals homozygous for the eSNP risk alleles (rs13003464-G and rs10181042-A), which are associated with CD. *C2orf74* is a gene of yet unknown function. In the Human Protein Atlas, *C2orf74* is expressed in all tissues, however, in the thymic APCs, *C2orf74* was not expressed according to our threshold (FPKM > 1), although low levels could detected in all four APCs. The SNP-probe pair 10181042-ILMN_339804 emerged as the strongest eQTL in the screen in **paper II**, and more studies should be performed to understand the functional role of *C2orf74* and the implication this eQTL can have in thymus.

5.1.9 AJ006998.2

Finally, the eSNP rs1736020-C risk allele associated with CD correlated with a lower expression level of AJ006998.2. This was the second potentially thymus-specific eQTL in **paper II**. AJ006998.2 is a lincRNA transcript, transcribed from the non-coding lincRNA DNA sequence AJ009632.2. AJ009632.2 did not seem to be expressed (FPKM > 1) in the thymic APCs, the exceptions were a single outlier in CD141+ and one in CD19+ B cells. However, it is not possible to know which transcript that is observed in these outliers, as the non-coding sequence AJ009632.2 is reported to have ten different splice variants according to the Ensembl database (http://www.ensembl.org). The general lack of AJ009632.2 expression in the thymic APCs suggests that the AJ006998.2 eQTL is a thymocyte eQTL. As mentioned in the introduction, the molecular function of lncRNAs is highly diverse [52], and lncRNAs are known to be expressed in both CD8+ T [173] cells and CD4+ T cells [174]. If AJ006998.2 is expressed in the thymocytes, the functional implication of this potential thymus-specific eQTL in T cell development could be an interesting focus for further studies.

5.2 Are the eGenes involved in any specific pathways?

A variety of databases can be used to analyse whether a set of genes are members of the same biological pathway. Pathway analysis with risk loci have been performed in several AIDs, including RA [175, 176], SLE [176], MS [177], PBC [178], T1D [179] and CD [179]. Interestingly, these studies led to the identification of several pathways that were shared among the various AIDs [23]. Since our eGenes were associated with risk variants from different AIDs (and therefore perhaps a common pathway for all) we analysed these genes using the ingenuity pathway (IPA) software. However, this did not result in any significant association to a specific pathway. The reason for this could be that the number of eGenes emerging from our eQTL screen (n = 8) was be too low, as biological pathways usually involve many genes. However, several of the shared pathways between immune-mediated diseases are immunological of nature [29, 44], which further supports to study the thymus organ in more detail.

5.3 Have we found all the AID-eQTLs?

In **paper II**, several of the SNP-Probe pairs in our screen which did not pass the significance thresholds (3.1×10^{-6} and 2.7×10^{-5}) could still be true eQTLs. As discussed in this paper, two eQTLs involving *IL18RAP* ($P = 1.8 \times 10^{-3}$) and *IRF5* ($P = 0.034$) have been reported in earlier studies [64, 65], but did not pass the strict significance thresholds in our screen. More AID-eQTLs could have emerged if we had a larger sample size and if we had taken account of certain aspects in the screen. As mentioned in Section 4.5 (Quality Control of the data) and 4.6 (Statistics), if we had tested only AID risk variants that were not linked by LD, or had removed probes that were either binding non-uniquely or overlapping with common SNPs (MAF > 1%) before the eQTL testing, fewer SNP-probe pair tests would have been included and the threshold of multiple testing would have been less strict. Other thymic eGenes from the screen in **paper II** with suggestive significance ($P < 7.33 \times 10^{-4}$) that have roles in inflammation and immune response included for instance *CLEC1* [180], *ELMO1* [181] and *TRAIP* [182]. Nevertheless, we chose to have a strict significance threshold, and rather accept more type II errors than type I errors in our screen.

5.4 The location of regulatory eSNPs relative to their eGenes

In the context of cis-eQTL mapping, is a window of 2 Mb (location of eGene ± 1 Mb) enough for investigating a sufficient number of SNPs and for finding the causal eSNP? Among the nine eQTLs that emerged from **paper I** and **II**, we observed that the lead eSNPs that correlated with expression of *ERAP1*, *FCRL3*, *SIRPG*, *RNASET2*, *SYS1* and *AJ006998.2* were positioned either within the eGene itself or upstream from the transcription start site of the eGene. The location of these eSNPs supports the theory that many cis-regulatory SNPs (>90%) are located within a window of ± 100 kb of the transcription start site [56, 60, 183]. The eSNPs that correlated with the expression levels of *ERAP2*, *C2orf74* and *NPIPB8* were positioned in one, three or five genes away from their respective eGenes, suggesting that cis-regulatory mechanisms can extend over longer distances, likely due transcriptional processes such as chromatin looping [184]. More precisely, these eSNPs were located 124,637 – 169,068 bp from their eGenes, hence not far from the 100 kb threshold. Although we did not have the power in our eQTL screens to pinpoint the causal lead eSNP, these findings suggest that a 1 Mb window is more than sufficient in order to find the causal peak eSNP for a gene.

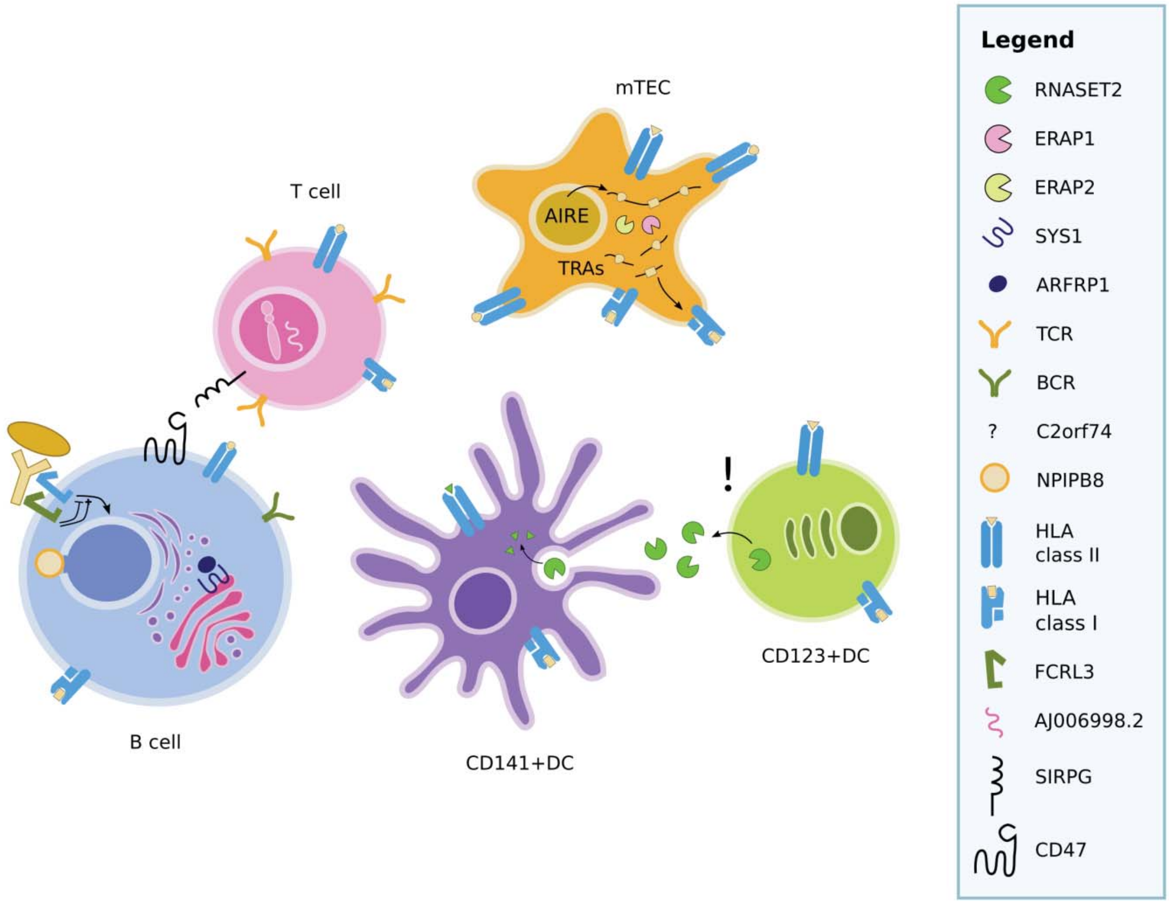

Figure 15: A Figure summarizing the eGene functions in the thymic APCs.

6 Conclusions and future perspectives

The studies performed here are some of the first to combine findings from GWAS and Immunochip studies with gene expression in thymus, and complement them with functional annotation from epigenome databases. These studies clearly indicate that thymic gene regulation might be influenced by autoimmune susceptibility loci. Some of the eQTLs were exclusively found in thymus, indicating that the disease risk SNPs possibly regulate gene expression uniquely in this organ. Nevertheless, it is highly likely that many AID-associated eQTLs from minor cell populations in thymus might have been lost due to the over-representation of developing thymocytes (95%). I believe that we have only started to uncover the tip of the iceberg, and that many more AID-related QTLs (eQTLs, splicing QTLs, transcription factor binding QTLs) would emerge from the transcriptome profiling of different thymic cell populations. The thymus is a complex tissue comprising many cell populations, and in the future we need to seize a higher resolution of the thymic cell subsets to determine which cell types that are affected by the risk SNPs.

In future studies, the main focus will be to increase sample size and to test novel AID risk variants for eQTL associations in thymus and in the different thymic cell populations. In order to achieve this, we need to dissect the thymic tissue by mapping all cell types (for instance by CyTOF), obtain higher purity of viable, single thymic cell types (by using a flow cytometry cell sorter, such as for example FACs Aria) and address their transcriptional landscape at different developmental stages. When more specialized technology has been developed, we need to assess the peptide repertoire on the HLA molecules of thymic APCs to confirm which TRAs are in fact presented to the developing thymocytes.

Collectively, these efforts will hopefully increase our understanding of the pathogenic mechanisms in AIDs, enabling better therapeutic options for patients. Advancement in the field of genetics in combination with understanding the functional role of these risk variants could lead to personalized medicine and novel therapeutic approaches that are based on particular autoimmune phenotypes and genomic alterations.

References

1. Cooper, G.S., M.L. Bynum, and E.C. Somers, *Recent insights in the epidemiology of autoimmune diseases: improved prevalence estimates and understanding of clustering of diseases.* J Autoimmun, 2009. **33**(3-4): p. 197-207.
2. Wang, L.F., F.S. Wang, and M.E. Gershwin, *Human autoimmune diseases: a comprehensive update.* Journal of Internal Medicine, 2015. **278**(4): p. 369-395.
3. Somers, E.C., et al., *Autoimmune diseases co-occurring within individuals and within families - A systematic review.* Epidemiology, 2006. **17**(2): p. 202-217.
4. Wang, L., F.S. Wang, and M.E. Gershwin, *Human autoimmune diseases: a comprehensive update.* J Intern Med, 2015. **278**(4): p. 369-95.
5. Cooper, G.S. and B.C. Stroehla, *The epidemiology of autoimmune diseases.* Autoimmunity Reviews, 2003. **2**(3): p. 119-125.
6. Toifl, K., J. Zeitlhofer, and B. Vass, *[The prognosis of juvenile myasthenia gravis].* Padiatr Padol, 1988. **23**(3): p. 195-207.
7. Walsh, S.J. and L.M. Rau, *Autoimmune diseases: A leading cause of death among young and middle-aged women in the United States.* American Journal of Public Health, 2000. **90**(9): p. 1463-1466.
8. Libert, C., L. Dejager, and I. Pinheiro, *The X chromosome in immune functions: when a chromosome makes the difference.* Nature Reviews Immunology, 2010. **10**(8): p. 594-604.
9. Bellone, M., *Autoimmune Disease: Pathogenesis.* Wiley Online Library, 2015.
10. Danke, N.A., et al., *Autoreactive T cells in healthy individuals.* J Immunol, 2004. **172**(10): p. 5967-72.
11. Koelsch, K., et al., *Mature B cells class switched to IgD are autoreactive in healthy individuals.* J Clin Invest, 2007. **117**(6): p. 1558-65.
12. Coutinho, A., M.D. Kazatchkine, and S. Avrameas, *Natural autoantibodies.* Current Opinion in Immunology, 1995. **7**(6): p. 812-818.
13. Palmer, J.P., et al., *Insulin-Antibodies in Insulin-Dependent Diabetics before Insulin-Treatment.* Science, 1983. **222**(4630): p. 1337-1339.
14. Bottazzo, G.F., A. Florin-Christensen, and D. Doniach, *Islet-cell antibodies in diabetes mellitus with autoimmune polyendocrine deficiencies.* Lancet, 1974. **2**(7892): p. 1279-83.
15. Bonifacio, E., V. Lampasona, and P.J. Bingley, *IA-2 (Islet cell antigen 512) is the primary target of humoral autoimmunity against type 1 diabetes-associated tyrosine phosphatase autoantigens.* Journal of Immunology, 1998. **161**(5): p. 2648-2654.
16. Baekkeskov, S., et al., *Identification of the 64k Autoantigen in Insulin-Dependent Diabetes as the Gaba-Synthesizing Enzyme Glutamic-Acid Decarboxylase.* Nature, 1990. **347**(6289): p. 151-156.
17. Riemekasten, G. and B.H. Hahn, *Key autoantigens in SLE.* Rheumatology, 2005. **44**(8): p. 975-982.
18. Chang, K., et al., *Smoking and rheumatoid arthritis.* Int J Mol Sci, 2014. **15**(12): p. 22279-95.
19. Dewar, D., S.P. Pereira, and P.J. Ciclitira, *The pathogenesis of coeliac disease.* Int J Biochem Cell Biol, 2004. **36**(1): p. 17-24.
20. Fradin, D., et al., *Association of the CpG methylation pattern of the proximal insulin gene promoter with type 1 diabetes.* PLoS One, 2012. **7**(5): p. e36278.
21. Calabrese, R., et al., *Methylation-dependent PAD2 upregulation in multiple sclerosis peripheral blood.* Mult Scler, 2012. **18**(3): p. 299-304.
22. Hu, N., et al., *Abnormal histone modification patterns in lupus CD4+ T cells.* J Rheumatol, 2008. **35**(5): p. 804-10.
23. Seldin, M.F., *The genetics of human autoimmune disease: A perspective on progress in the field and future directions.* J Autoimmun, 2015. **64**: p. 1-12.

24. Ezkurdia, I., et al., *Multiple evidence strands suggest that there may be as few as 19 000 human protein-coding genes.* Human Molecular Genetics, 2014. **23**(22): p. 5866-5878.
25. Gonzaga-Jauregui, C., J.R. Lupski, and R.A. Gibbs, *Human genome sequencing in health and disease.* Annu Rev Med, 2012. **63**: p. 35-61.
26. Gibbs, R.A., et al., *The International HapMap Project.* Nature, 2003. **426**(6968): p. 789-796.
27. Zhang, W.Q., et al., *Comparing genetic variants detected in the 1000 genomes project with SNPs determined by the International HapMap Consortium.* Journal of Genetics, 2015. **94**(4): p. 731-740.
28. Dudbridge, F. and A. Gusnanto, *Estimation of significance thresholds for genomewide association scans.* Genetic Epidemiology, 2008. **32**(3): p. 227-234.
29. Parkes, M., et al., *Genetic insights into common pathways and complex relationships among immune-mediated diseases.* Nature Reviews Genetics, 2013. **14**(9): p. 661-673.
30. Cortes, A. and M.A. Brown, *Promise and pitfalls of the Immunochip.* Arthritis Res Ther, 2011. **13**(1): p. 101.
31. International Genetics of Ankylosing Spondylitis, C., et al., *Identification of multiple risk variants for ankylosing spondylitis through high-density genotyping of immune-related loci.* Nat Genet, 2013. **45**(7): p. 730-8.
32. Eyre, S., et al., *High-density genetic mapping identifies new susceptibility loci for rheumatoid arthritis.* Nat Genet, 2012. **44**(12): p. 1336-40.
33. Cooper, J.D., et al., *Seven newly identified loci for autoimmune thyroid disease.* Hum Mol Genet, 2012. **21**(23): p. 5202-8.
34. Tsoi, L.C., et al., *Identification of 15 new psoriasis susceptibility loci highlights the role of innate immunity.* Nat Genet, 2012. **44**(12): p. 1341-8.
35. Trynka, G., et al., *Dense genotyping identifies and localizes multiple common and rare variant association signals in celiac disease.* Nat Genet, 2011. **43**(12): p. 1193-201.
36. Jostins, L., et al., *Host-microbe interactions have shaped the genetic architecture of inflammatory bowel disease.* Nature, 2012. **491**(7422): p. 119-24.
37. International Multiple Sclerosis Genetics, C., et al., *Analysis of immune-related loci identifies 48 new susceptibility variants for multiple sclerosis.* Nat Genet, 2013. **45**(11): p. 1353-60.
38. Ellinghaus, D., et al., *High-density genotyping study identifies four new susceptibility loci for atopic dermatitis.* Nature Genetics, 2013. **45**(7): p. 808-+.
39. Liu, J.Z., et al., *Dense genotyping of immune-related disease regions identifies nine new risk loci for primary sclerosing cholangitis.* Nat Genet, 2013. **45**(6): p. 670-5.
40. Mayes, M.D., et al., *Immunochip Analysis Identifies Multiple Susceptibility Loci for Systemic Sclerosis.* American Journal of Human Genetics, 2014. **94**(1): p. 47-61.
41. Hinks, A., et al., *Dense genotyping of immune-related disease regions identifies 14 new susceptibility loci for juvenile idiopathic arthritis.* Nat Genet, 2013. **45**(6): p. 664-9.
42. Onengut-Gumuscu, S., et al., *Fine mapping of type 1 diabetes susceptibility loci and evidence for colocalization of causal variants with lymphoid gene enhancers.* Nat Genet, 2015. **47**(4): p. 381-6.
43. Jonkers, I.H. and C. Wijmenga, *Context-specific effects of genetic variants associated with autoimmune disease.* Hum Mol Genet, 2017. **26**(R2): p. R185-R192.
44. Zhernakova, A., C.C. van Diemen, and C. Wijmenga, *Detecting shared pathogenesis from the shared genetics of immune-related diseases.* Nature Reviews Genetics, 2009. **10**(1): p. 43-55.
45. Okada, Y., et al., *Genetics of rheumatoid arthritis contributes to biology and drug discovery.* Nature, 2014. **506**(7488): p. 376-81.
46. Ricano-Ponce, I. and C. Wijmenga, *Mapping of immune-mediated disease genes.* Annu Rev Genomics Hum Genet, 2013. **14**: p. 325-53.
47. Hu, X., et al., *Integrating autoimmune risk loci with gene-expression data identifies specific pathogenic immune cell subsets.* Am J Hum Genet, 2011. **89**(4): p. 496-506.
48. Zhang, W., et al., *Effector CD4+ T cell expression signatures and immune-mediated disease associated genes.* PLoS One, 2012. **7**(6): p. e38510.

49. Myers, R.M., et al., *A User's Guide to the Encyclopedia of DNA Elements (ENCODE).* Plos Biology, 2011. **9**(4).
50. Chadwick, L.H., *The NIH Roadmap Epigenomics Program data resource.* Epigenomics, 2012. **4**(3): p. 317-24.
51. Roadmap Epigenomics, C., et al., *Integrative analysis of 111 reference human epigenomes.* Nature, 2015. **518**(7539): p. 317-30.
52. Fitzgerald, K.A. and D.R. Caffrey, *Long noncoding RNAs in innate and adaptive immunity.* Curr Opin Immunol, 2014. **26**: p. 140-6.
53. Kumar, V., et al., *Human disease-associated genetic variation impacts large intergenic non-coding RNA expression.* PLoS Genet, 2013. **9**(1): p. e1003201.
54. Stranger, B.E. and T. Raj, *Genetics of human gene expression.* Curr Opin Genet Dev, 2013. **23**(6): p. 627-34.
55. Fu, J., et al., *Unraveling the regulatory mechanisms underlying tissue-dependent genetic variation of gene expression.* PLoS Genet, 2012. **8**(1): p. e1002431.
56. Stranger, B.E., et al., *Population genomics of human gene expression.* Nat Genet, 2007. **39**(10): p. 1217-24.
57. Zhong, H., et al., *Liver and adipose expression associated SNPs are enriched for association to type 2 diabetes.* PLoS Genet, 2010. **6**(5): p. e1000932.
58. Heap, G.A., et al., *Complex nature of SNP genotype effects on gene expression in primary human leucocytes.* BMC Med Genomics, 2009. **2**: p. 1.
59. Myers, A.J., et al., *A survey of genetic human cortical gene expression.* Nat Genet, 2007. **39**(12): p. 1494-9.
60. Emilsson, V., et al., *Genetics of gene expression and its effect on disease.* Nature, 2008. **452**(7186): p. 423-8.
61. Ding, J., et al., *Gene expression in skin and lymphoblastoid cells: Refined statistical method reveals extensive overlap in cis-eQTL signals.* Am J Hum Genet, 2010. **87**(6): p. 779-89.
62. Ward, L.D. and M. Kellis, *HaploReg: a resource for exploring chromatin states, conservation, and regulatory motif alterations within sets of genetically linked variants.* Nucleic Acids Research, 2012. **40**(D1): p. D930-D934.
63. Ward, L.D. and M. Kellis, *HaploReg v4: systematic mining of putative causal variants, cell types, regulators and target genes for human complex traits and disease.* Nucleic Acids Res, 2016. **44**(D1): p. D877-81.
64. Amundsen, S.S., et al., *Coeliac disease-associated polymorphisms influence thymic gene expression.* Genes Immun, 2014.
65. Nordang, G.B.N., et al., *Interferon regulatory factor 5 gene polymorphism confers risk to several rheumatic diseases and correlates with expression of alternative thymic transcripts.* Rheumatology, 2012. **51**(4): p. 619-626.
66. Mero, I.L., et al., *Exploring the CLEC16A gene reveals a MS-associated variant with correlation to the relative expression of CLEC16A isoforms in thymus.* Genes Immun, 2011. **12**(3): p. 191-8.
67. Kyewski, B. and L. Klein, *A central role for central tolerance.* Annu Rev Immunol, 2006. **24**: p. 571-606.
68. Huesmann, M., et al., *Kinetics and efficacy of positive selection in the thymus of normal and T cell receptor transgenic mice.* Cell, 1991. **66**(3): p. 533-40.
69. Shortman, K., D. Vremec, and M. Egerton, *The kinetics of T cell antigen receptor expression by subgroups of CD4+8+ thymocytes: delineation of CD4+8+3(2+) thymocytes as post-selection intermediates leading to mature T cells.* J Exp Med, 1991. **173**(2): p. 323-32.
70. Ignatowicz, L., J. Kappler, and P. Marrack, *The repertoire of T cells shaped by a single MHC/peptide ligand.* Cell, 1996. **84**(4): p. 521-9.
71. Xing, Y. and K.A. Hogquist, *T-cell tolerance: central and peripheral.* Cold Spring Harb Perspect Biol, 2012. **4**(6).

72. Klein, L., et al., *Positive and negative selection of the T cell repertoire: what thymocytes see (and don't see).* Nat Rev Immunol, 2014. **14**(6): p. 377-91.
73. Derbinski, J., et al., *Promiscuous gene expression in medullary thymic epithelial cells mirrors the peripheral self.* Nat Immunol, 2001. **2**(11): p. 1032-9.
74. Sansom, S.N., et al., *Population and single-cell genomics reveal the Aire dependency, relief from Polycomb silencing, and distribution of self-antigen expression in thymic epithelia.* Genome Res, 2014. **24**(12): p. 1918-31.
75. Pinto, S., et al., *Overlapping gene coexpression patterns in human medullary thymic epithelial cells generate self-antigen diversity.* Proc Natl Acad Sci U S A, 2013. **110**(37): p. E3497-505.
76. St-Pierre, C., et al., *Differential Features of AIRE-Induced and AIRE-Independent Promiscuous Gene Expression in Thymic Epithelial Cells.* J Immunol, 2015. **195**(2): p. 498-506.
77. Org, T., et al., *The autoimmune regulator PHD finger binds to non-methylated histone H3K4 to activate gene expression.* EMBO Rep, 2008. **9**(4): p. 370-6.
78. Oven, I., et al., *AIRE recruits P-TEFb for transcriptional elongation of target genes in medullary thymic epithelial cells.* Mol Cell Biol, 2007. **27**(24): p. 8815-23.
79. Anderson, M.S. and M.A. Su, *AIRE expands: new roles in immune tolerance and beyond.* Nat Rev Immunol, 2016. **16**(4): p. 247-58.
80. Giraud, M., et al., *Aire unleashes stalled RNA polymerase to induce ectopic gene expression in thymic epithelial cells.* Proc Natl Acad Sci U S A, 2012. **109**(2): p. 535-40.
81. Klein, L., et al., *Antigen presentation in the thymus for positive selection and central tolerance induction.* Nat Rev Immunol, 2009. **9**(12): p. 833-44.
82. Gurka, S., et al., *Expression analysis of surface molecules on human thymic dendritic cells with the 10th HLDA Workshop antibody panel.* Clin Transl Immunology, 2015. **4**(10): p. e47.
83. Vandenabeele, S., et al., *Human thymus contains 2 distinct dendritic cell populations.* Blood, 2001. **97**(6): p. 1733-41.
84. Baba, T., Y. Nakamoto, and N. Mukaida, *Crucial contribution of thymic Sirp alpha+ conventional dendritic cells to central tolerance against blood-borne antigens in a CCR2-dependent manner.* J Immunol, 2009. **183**(5): p. 3053-63.
85. Hadeiba, H., et al., *Plasmacytoid dendritic cells transport peripheral antigens to the thymus to promote central tolerance.* Immunity, 2012. **36**(3): p. 438-50.
86. Moore, N.C., et al., *Differential Expression of Mtv Loci in Mhc Class Ii-Positive Thymic Stromal Cells.* Journal of Immunology, 1994. **152**(10): p. 4826-4831.
87. Perera, J., et al., *Autoreactive thymic B cells are efficient antigen-presenting cells of cognate self-antigens for T cell negative selection.* Proc Natl Acad Sci U S A, 2013. **110**(42): p. 17011-6.
88. Fujihara, C., et al., *T cell-B cell thymic cross-talk: maintenance and function of thymic B cells requires cognate CD40-CD40 ligand interaction.* J Immunol, 2014. **193**(11): p. 5534-44.
89. Yamano, T., et al., *Thymic B Cells Are Licensed to Present Self Antigens for Central T Cell Tolerance Induction.* Immunity, 2015. **42**(6): p. 1048-61.
90. Frommer, F. and A. Waisman, *B cells participate in thymic negative selection of murine auto-reactive CD4+ T cells.* PLoS One, 2010. **5**(10): p. e15372.
91. Cepeda, S., et al., *Age-Associated Decline in Thymic B Cell Expression of Aire and Aire-Dependent Self-Antigens.* Cell Rep, 2018. **22**(5): p. 1276-1287.
92. Gies, V., et al., *B cells differentiate in human thymus and express AIRE.* J Allergy Clin Immunol, 2017. **139**(3): p. 1049-1052 e12.
93. Nicolae, D.L., et al., *Trait-Associated SNPs Are More Likely to Be eQTLs: Annotation to Enhance Discovery from GWAS.* Plos Genetics, 2010. **6**(4).
94. Evans, D.M., et al., *Interaction between ERAP1 and HLA-B27 in ankylosing spondylitis implicates peptide handling in the mechanism for HLA-B27 in disease susceptibility.* Nature Genetics, 2011. **43**(8): p. 761-U67.
95. Australo-Anglo-American Spondyloarthritis, C., et al., *Genome-wide association study of ankylosing spondylitis identifies non-MHC susceptibility loci.* Nat Genet, 2010. **42**(2): p. 123-7.

96. Franke, A., et al., *Genome-wide meta-analysis increases to 71 the number of confirmed Crohn's disease susceptibility loci.* Nat Genet, 2010. **42**(12): p. 1118-25.

97. Guerini, F.R., et al., *A functional variant in ERAP1 predisposes to multiple sclerosis.* PLoS One, 2012. **7**(1): p. e29931.

98. Genetic Analysis of Psoriasis, C., et al., *A genome-wide association study identifies new psoriasis susceptibility loci and an interaction between HLA-C and ERAP1.* Nat Genet, 2010. **42**(11): p. 985-90.

99. Fung, E.Y., et al., *Analysis of 17 autoimmune disease-associated variants in type 1 diabetes identifies 6q23/TNFAIP3 as a susceptibility locus.* Genes Immun, 2009. **10**(2): p. 188-91.

100. Dendrou, C.A., et al., *Cell-specific protein phenotypes for the autoimmune locus IL2RA using a genotype-selectable human bioresource.* Nature Genetics, 2009. **41**(9): p. 1011-U80.

101. Vremec, D., *The isolation of mouse dendritic cells from lymphoid tissues and the identification of dendritic cell subtypes by multiparameter flow cytometry.* Methods Mol Biol, 2010. **595**: p. 205-29.

102. Stoeckle, C., et al., *Isolation of myeloid dendritic cells and epithelial cells from human thymus.* J Vis Exp, 2013(79): p. e50951.

103. Gray, D.H., A.P. Chidgey, and R.L. Boyd, *Analysis of thymic stromal cell populations using flow cytometry.* J Immunol Methods, 2002. **260**(1-2): p. 15-28.

104. Seach, N., et al., *Purified enzymes improve isolation and characterization of the adult thymic epithelium.* Journal of Immunological Methods, 2012. **385**(1-2): p. 23-34.

105. Gray, D.H., et al., *Unbiased analysis, enrichment and purification of thymic stromal cells.* J Immunol Methods, 2008. **329**(1-2): p. 56-66.

106. Januszyk, M., et al., *Evaluating the Effect of Cell Culture on Gene Expression in Primary Tissue Samples Using Microfluidic-Based Single Cell Transcriptional Analysis.* Microarrays (Basel), 2015. **4**(4): p. 540-50.

107. Liu, Y., J. Zhou, and K.P. White, *RNA-seq differential expression studies: more sequence or more replication?* Bioinformatics, 2014. **30**(3): p. 301-4.

108. Chomczynski, P., *A reagent for the single-step simultaneous isolation of RNA, DNA and proteins from cell and tissue samples.* Biotechniques, 1993. **15**(3): p. 532-4, 536-7.

109. Miragaia, R.J., et al., *Single-cell RNA-sequencing resolves self-antigen expression during mTEC development.* Sci Rep, 2018. **8**(1): p. 685.

110. Barrett, J.C., et al., *Genome-wide association defines more than 30 distinct susceptibility loci for Crohn's disease.* Nature Genetics, 2008. **40**(8): p. 955-962.

111. Burton, P.R., et al., *Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls.* Nature, 2007. **447**(7145): p. 661-678.

112. Dubois, P.C., et al., *Multiple common variants for celiac disease influencing immune gene expression.* Nat Genet, 2010. **42**(4): p. 295-302.

113. Kugathasan, S., et al., *Loci on 20q13 and 21q22 are associated with pediatric-onset inflammatory bowel disease.* Nat Genet, 2008. **40**(10): p. 1211-5.

114. Duerr, R.H., et al., *A genome-wide association study identifies IL23R as an inflammatory bowel disease gene.* Science, 2006. **314**(5804): p. 1461-1463.

115. International Multiple Sclerosis Genetics, C., et al., *Genetic risk and a primary role for cell-mediated immune mechanisms in multiple sclerosis.* Nature, 2011. **476**(7359): p. 214-9.

116. Patsopoulos, N.A., et al., *Genome-wide meta-analysis identifies novel multiple sclerosis susceptibility loci.* Annals of Neurology, 2011. **70**(6): p. 897-912.

117. De Jager, P.L., et al., *Meta-analysis of genome scans and replication identify CD6, IRF8 and TNFRSF1A as new multiple sclerosis susceptibility loci.* Nature Genetics, 2009. **41**(7): p. 776-U26.

118. Baranzini, S.E., et al., *Genome-wide association analysis of susceptibility and clinical phenotype in multiple sclerosis.* Hum Mol Genet, 2009. **18**(4): p. 767-78.

119. International Multiple Sclerosis Genetics, C., et al., *Risk alleles for multiple sclerosis identified by a genomewide study.* N Engl J Med, 2007. **357**(9): p. 851-62.

120. Comabella, M., et al., *Identification of a Novel Risk Locus for Multiple Sclerosis at 13q31.3 by a Pooled Genome-Wide Scan of 500,000 Single Nucleotide Polymorphisms.* Plos One, 2008. **3**(10).
121. Aulchenko, Y.S., et al., *Genetic variation in the KIF1B locus influences susceptibility to multiple sclerosis.* Nature Genetics, 2008. **40**(12): p. 1402-1403.
122. Jakkula, E., et al., *Genome-wide Association Study in a High-Risk Isolate for Multiple Sclerosis Reveals Associated Variants in STAT3 Gene.* American Journal of Human Genetics, 2010. **86**(2): p. 285-291.
123. Sanna, S., et al., *Variants within the immunoregulatory CBLB gene are associated with multiple sclerosis.* Nature Genetics, 2010. **42**(6): p. 495-497.
124. Bahlo, M., et al., *Genome-wide association study identifies new multiple sclerosis susceptibility loci on chromosomes 12 and 20.* Nature Genetics, 2009. **41**(7): p. 824-U84.
125. Wang, J.H., et al., *Modeling the cumulative genetic risk for multiple sclerosis from genome-wide association data.* Genome Medicine, 2011. **3**.
126. Nischwitz, S., et al., *Evidence for VAV2 and ZNF433 as susceptibility genes for multiple sclerosis.* Journal of Neuroimmunology, 2010. **227**(1-2): p. 162-166.
127. Briggs, F.B.S., et al., *Genome-wide association study of severity in multiple sclerosis.* Genes and Immunity, 2011. **12**(8): p. 615-625.
128. Nair, R.P., et al., *Genome-wide scan reveals association of psoriasis with IL-23 and NF-kappaB pathways.* Nat Genet, 2009. **41**(2): p. 199-204.
129. Stuart, P.E., et al., *Genome-wide association analysis identifies three psoriasis susceptibility loci.* Nature Genetics, 2010. **42**(11): p. 1000-U125.
130. Liu, Y., et al., *A genome-wide association study of psoriasis and psoriatic arthritis identifies new disease loci.* Plos Genetics, 2008. **4**(4).
131. Stahl, E.A., et al., *Genome-wide association study meta-analysis identifies seven new rheumatoid arthritis risk loci.* Nat Genet, 2010. **42**(6): p. 508-14.
132. Kozyrev, S.V., et al., *Functional variants in the B-cell gene BANK1 are associated with systemic lupus erythematosus.* Nat Genet, 2008. **40**(2): p. 211-6.
133. Graham, R.R., et al., *Genetic variants near TNFAIP3 on 6q23 are associated with systemic lupus erythematosus.* Nat Genet, 2008. **40**(9): p. 1059-61.
134. International Consortium for Systemic Lupus Erythematosus, G., et al., *Genome-wide association scan in women with systemic lupus erythematosus identifies susceptibility variants in ITGAM, PXK, KIAA1542 and other loci.* Nat Genet, 2008. **40**(2): p. 204-10.
135. Hom, G., et al., *Association of systemic lupus erythematosus with C8orf13-BLK and ITGAM-ITGAX.* N Engl J Med, 2008. **358**(9): p. 900-9.
136. Barrett, J.C., et al., *Genome-wide association study and meta-analysis find that over 40 loci affect risk of type 1 diabetes.* Nature Genetics, 2009. **41**(6): p. 703-707.
137. Huang, J., et al., *1000 Genomes-based imputation identifies novel and refined associations for the Wellcome Trust Case Control Consortium phase 1 Data.* European Journal of Human Genetics, 2012. **20**(7): p. 801-805.
138. Bradfield, J.P., et al., *A Genome-Wide Meta-Analysis of Six Type 1 Diabetes Cohorts Identifies Multiple Associated Loci.* Plos Genetics, 2011. **7**(9).
139. Anderson, C.A., et al., *Meta-analysis identifies 29 additional ulcerative colitis risk loci, increasing the number of confirmed associations to 47.* Nature Genetics, 2011. **43**(3): p. 246-U94.
140. Wang, Z., M. Gerstein, and M. Snyder, *RNA-Seq: a revolutionary tool for transcriptomics.* Nat Rev Genet, 2009. **10**(1): p. 57-63.
141. Williams, A.G., et al., *RNA-seq Data: Challenges in and Recommendations for Experimental Design and Analysis.* Curr Protoc Hum Genet, 2014. **83**: p. 11 13 1-20.
142. Finotello, F. and B. Di Camillo, *Measuring differential gene expression with RNA-seq: challenges and strategies for data analysis.* Brief Funct Genomics, 2015. **14**(2): p. 130-42.

143. Bhargava, V., et al., *Technical variations in low-input RNA-seq methodologies.* Sci Rep, 2014. **4**: p. 3678.

144. Ramskold, D., et al., *Full-length mRNA-Seq from single-cell levels of RNA and individual circulating tumor cells.* Nat Biotechnol, 2012. **30**(8): p. 777-82.

145. Bhargava, V., et al., *Quantitative transcriptomics using designed primer-based amplification.* Sci Rep, 2013. **3**: p. 1740.

146. Dobin, A., et al., *STAR: ultrafast universal RNA-seq aligner.* Bioinformatics, 2013. **29**(1): p. 15-21.

147. Dobin, A. and T.R. Gingeras, *Mapping RNA-seq Reads with STAR.* Curr Protoc Bioinformatics, 2015. **51**: p. 11 14 1-19.

148. Dozmorov, M.G., et al., *Detrimental effects of duplicate reads and low complexity regions on RNA- and ChIP-seq data.* BMC Bioinformatics, 2015. **16 Suppl 13**: p. S10.

149. Parekh, S., et al., *The impact of amplification on differential expression analyses by RNA-seq.* Sci Rep, 2016. **6**: p. 25533.

150. Robinson, M.D., D.J. McCarthy, and G.K. Smyth, *edgeR: a Bioconductor package for differential expression analysis of digital gene expression data.* Bioinformatics, 2010. **26**(1): p. 139-140.

151. Alcina, A., et al., *Multiple sclerosis risk variant HLA-DRB1*1501 associates with high expression of DRB1 gene in different human populations.* PLoS One, 2012. **7**(1): p. e29819.

152. Ramasamy, A., et al., *Resolving the polymorphism-in-probe problem is critical for correct interpretation of expression QTL studies.* Nucleic Acids Res, 2013. **41**(7): p. e88.

153. Purcell, S., et al., *PLINK: a tool set for whole-genome association and population-based linkage analyses.* Am J Hum Genet, 2007. **81**(3): p. 559-75.

154. Uhlen, M., et al., *Proteomics. Tissue-based map of the human proteome.* Science, 2015. **347**(6220): p. 1260419.

155. Adamopoulou, E., et al., *Exploring the MHC-peptide matrix of central tolerance in the human thymus.* Nat Commun, 2013. **4**: p. 2039.

156. Alvarez, I., et al., *Central T cell tolerance: Identification of tissue-restricted autoantigens in the thymus HLA-DR peptidome.* J Autoimmun, 2015. **60**: p. 12-9.

157. Fierabracci, A., et al., *The putative role of endoplasmic reticulum aminopeptidases in autoimmunity: insights from genomic-wide association studies.* Autoimmun Rev, 2012. **12**(2): p. 281-8.

158. Andres, A.M., et al., *Balancing Selection Maintains a Form of ERAP2 that Undergoes Nonsense-Mediated Decay and Affects Antigen Presentation.* Plos Genetics, 2010. **6**(10).

159. Chen, Y., et al., *edgeR : differential expression analysis of digital gene expression data User ' s Guide.* 2015.

160. Napolitano, G., et al., *High glucose levels increase major histocompatibility complex class I gene expression in thyroid cells and amplify interferon-gamma action.* Endocrinology, 2002. **143**(3): p. 1008-17.

161. Serreze, D.V., et al., *Major histocompatibility complex class I-deficient NOD-B2mnull mice are diabetes and insulitis resistant.* Diabetes, 1994. **43**(3): p. 505-9.

162. Barclay, A.N. and M.H. Brown, *The SIRP family of receptors and immune regulation.* Nature Reviews Immunology, 2006. **6**(6): p. 457-464.

163. Piccio, L., et al., *Adhesion of human T cells to antigen-presenting cells through SIRPbeta2-CD47 interaction costimulates T-cell proliferation.* Blood, 2005. **105**(6): p. 2421-7.

164. Brooke, G., et al., *Human lymphocytes interact directly with CD47 through a novel member of the signal regulatory protein (SIRP) family.* J Immunol, 2004. **173**(4): p. 2562-70.

165. Davis, R.S., et al., *Identification of a family of Fc receptor homologs with preferential B cell expression.* Proceedings of the National Academy of Sciences of the United States of America, 2001. **98**(17): p. 9772-9777.

166. Jin, Y., et al., *Genome-wide association studies of autoimmune vitiligo identify 23 new risk loci and highlight key pathways and regulatory variants.* Nat Genet, 2016. **48**(11): p. 1418-1424.

167. Campomenosi, P., et al., *Characterization of RNASET2, the first human member of the Rh/T2/S family of glycoproteins.* Arch Biochem Biophys, 2006. **449**(1-2): p. 17-26.
168. Wang, Q., X. Wang, and L. Xiang, *Role and Mechanism of RNASET2 in the Pathogenesis of Vitiligo.* J Investig Dermatol Symp Proc, 2015. **17**(1): p. 48-50.
169. Everts, B., et al., *Schistosome-derived omega-1 drives Th2 polarization by suppressing protein synthesis following internalization by the mannose receptor.* J Exp Med, 2012. **209**(10): p. 1753-67, S1.
170. Behnia, R., et al., *Targeting of the Arf-like GTPase Arl3p to the Golgi requires N-terminal acetylation and the membrane protein Sys1p.* Nat Cell Biol, 2004. **6**(5): p. 405-13.
171. Graham, T.R., *Membrane targeting: getting Arl to the Golgi.* Curr Biol, 2004. **14**(12): p. R483-5.
172. Blum, J.S., P.A. Wearsch, and P. Cresswell, *Pathways of antigen processing.* Annu Rev Immunol, 2013. **31**: p. 443-73.
173. Pang, K.C., et al., *Genome-wide identification of long noncoding RNAs in CD8+ T cells.* J Immunol, 2009. **182**(12): p. 7738-48.
174. Pagani, M., et al., *Role of microRNAs and long-non-coding RNAs in CD4(+) T-cell differentiation.* Immunol Rev, 2013. **253**(1): p. 82-96.
175. Zhang, M.M., et al., *Pathway-based association analysis of two genome-wide screening data identifies rheumatoid arthritis-related pathways.* Genes Immun, 2014. **15**(7): p. 487-94.
176. Lee, Y.H., et al., *Genome-wide pathway analysis of genome-wide association studies on systemic lupus erythematosus and rheumatoid arthritis.* Mol Biol Rep, 2012. **39**(12): p. 10627-35.
177. International Multiple Sclerosis Genetics, C., *Network-based multiple sclerosis pathway analysis with GWAS data from 15,000 cases and 30,000 controls.* Am J Hum Genet, 2013. **92**(6): p. 854-65.
178. Kar, S.P., et al., *Pathway-based analysis of primary biliary cirrhosis genome-wide association studies.* Genes Immun, 2013. **14**(3): p. 179-86.
179. Carbonetto, P. and M. Stephens, *Integrated enrichment analysis of variants and pathways in genome-wide association studies indicates central role for IL-2 signaling genes in type 1 diabetes, and cytokine signaling genes in Crohn's disease.* PLoS Genet, 2013. **9**(10): p. e1003770.
180. Thebault, P., et al., *The C-type lectin-like receptor CLEC-1, expressed by myeloid cells and endothelial cells, is up-regulated by immunoregulatory mediators and moderates T cell activation.* J Immunol, 2009. **183**(5): p. 3099-108.
181. Das, S., et al., *ELMO1 has an essential role in the internalization of Salmonella Typhimurium into enteric macrophages that impacts disease outcome.* Cell Mol Gastroenterol Hepatol, 2015. **1**(3): p. 311-324.
182. Kong, Q.Z., et al., *Anti-Inflammatory Effects of TRAF-Interacting Protein in Rheumatoid Arthritis Fibroblast-Like Synoviocytes.* Mediators Inflamm, 2016. **2016**: p. 3906108.
183. Murphy, A., et al., *Mapping of numerous disease-associated expression polymorphisms in primary peripheral blood CD4+ lymphocytes.* Hum Mol Genet, 2010. **19**(23): p. 4745-57.
184. Krivega, I. and A. Dean, *Enhancer and promoter interactions-long distance calls.* Curr Opin Genet Dev, 2012. **22**(2): p. 79-85.

ERRATA

The following changes were made after submission of the thesis, but before printing:

Page 23 (Paragraph 1.5.1): «Any interference with the thymic medulla will manifest in autoimmunity, whether it is disruption of the three-dimentional space, …" was changed to "Any interference with the thymic medulla will manifest in autoimmunity, whether it is disruption of the three-**dimensional** space, …)"

Paper III:

Page 10 (Results, under "Genes encoding tissue-enriched proteins in the thymic APCs")

"Therefore, we reanalyzed the data with a lower threshold, where genes only needed to be present i one biological replicate…" was changed to "Therefore, we reanalyzed the data with a lower threshold, where genes only needed to be present **in** one biological replicate"

Page 10 (Results, under "Genes encoding tissue-enriched proteins in the thymic APCs")

"Finally, we investigated whether there were any overlap between the TRA genes and human autoantigens from the litearture (Fig 4B)." was changed to "Finally, we investigated whether there were any overlap between the TRA genes and human autoantigens from the **literature** (Fig 4B)."

Page 13 (Discussion):

"Furthermore, Liu et al. also reports that, for DE studies, sequencing more than 10 million reads per sample gives dimishing returns compared with adding replication (Liu et al. 2014)" was changed to "Furthermore, Liu et al. also reports that, for DE studies, sequencing more than 10 million reads per sample gives **diminishing** returns compared with adding replication (Liu et al. 2014)"

Page 16 (Methods, under "Isolation of thymic APCs")

"… before TECs were EpCam-positively selected with CELLection TM Epithelial Enrich (Thermo Fischer #16203)." was changed to "… before TECs were **EpCAM**-positively selected with CELLection TM Epithelial Enrich (Thermo Fischer #16203)."

Page 19 (Methods, under "Transcriptional regulator genes and genes encoding autoantigens in the thymic APCs")

"However, as TRAs are known to be lowly expressed in only 1-3% of mTECs at any given time, we lowered the filtering criterias in edgeR for this analysis and included genes present in at least one biological replicate in the dataset." was changed to "However, as TRAs are known to be lowly expressed in only 1-3% of mTECs at any given time, we lowered the filtering **criteria** in edgeR for this analysis and included genes present in at least one biological replicate in the dataset."

Page 20 (Acknowledgements):

"Aknowledgements" was changed to "**Acknowledgements**"

Supplementary Figure S1:

"Genes encoding protein markers in A. mTECs (EpCam, FOXN1 and AIRE) and in B. CD19+ B cells (CD19, CD22 and CD20 (MS4A1))." was changed to "Genes encoding protein markers in A. mTECs (**EpCAM**, FOXN1 and AIRE) and in B. CD19+ B cells (CD19, CD22 and CD20 (MS4A1))."