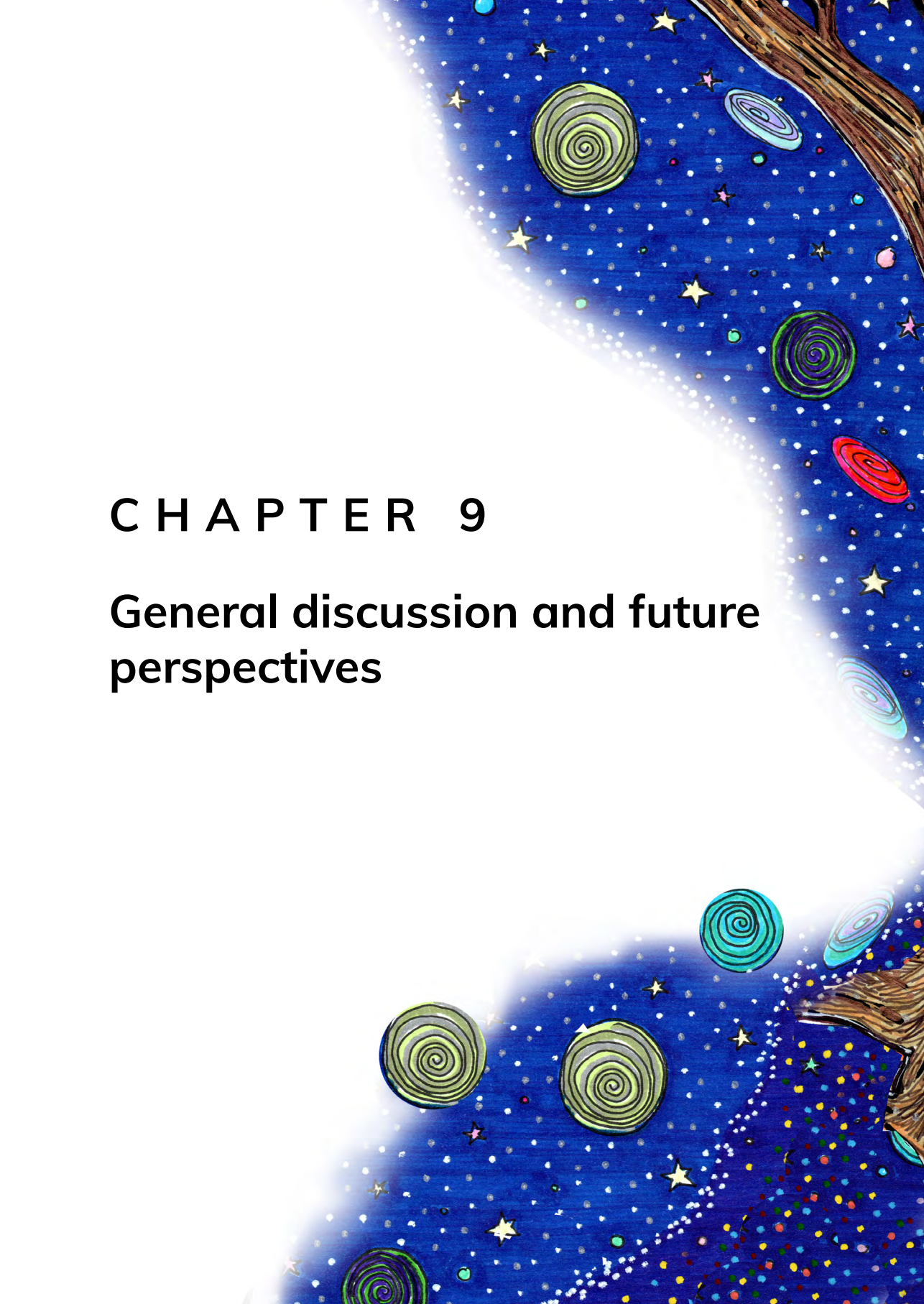
Towards finding and understanding the missing heritability of immune-mediated diseases

Ricaño Ponce, Isis



CHAPTER 9

General discussion and future perspectives

Where are we now?

Ten years ago the immunogenetics field was focused on discovering the loci that contribute to the risk of immune-mediated diseases (IMD). This search has been very successful, with almost fifty thousand loci so far implicated in multiple complex diseases[1]. For IMD alone, more than 300 loci have been described (Table 1) and replicated in various populations using different genotyping platforms. Intriguingly, the majority of associated loci contain mainly common variants (minor allele frequency (MAF) >10%) that each confer only small risk of developing the disease. For example, with the exception of the MHC region[2], the odds ratios of these IMD-associated variants vary from 1.04 to 3.99 (mean = 1.29), explaining only a small proportion of the heritability. With this data in hand the question became how to explain this missing heritability. I started working on this thesis at this point and this is reflected in the second part of my thesis, which focuses primarily on identifying additional genetic factors that contribute to celiac disease (CeD) heritability. However, over the years of my PhD, and after the identification of many loci, the main aim of the field shifted toward understanding the effect of the associated loci and pinpointing the true causal variants and genes in order to gain insight into the disease biology. Thus, in the first part of my thesis I characterized loci associated to IMD and focussed on refining the genetic associations of IMD identified by genome-wide association studies (GWAS) and Immunochip.

Although the Immunochip, a single nucleotide polymorphism (SNP) array, has proven useful for discovering new IMD loci and reducing the size of previously associated regions, pinpointing the true causal variants and genes has remained a challenge. As most associated regions contain multiple SNPs in high linkage disequilibrium (LD) and multiple genes, it was difficult to establish a causal link between IMD-associated SNPs and genes. Another challenge was understanding the associations coming from GWAS and Immunochip analysis where, as shown in chapter one, 90% of the associated variants are in the non-coding part of the genome.

Table 1. Immune-mediated diseases associated loci by GWAS and Immunochip

Disease	Abbreviation	GWAS	All associated loci*
Atopic dermatitis	AD	4	11
Ankylosing spondylitis	AS	11	23
Autoimmune thyroid disease	ATD	5 (Grave's disease)	8
Crohn's disease	CD	70	122
Celiac disease	CeD	26	41
Immunoglobulin A deficiency	IgAD	1	NA
Juvenile idiopathic arthritis	JIA	3	23
Multiple sclerosis	MS	41	106
Primary biliary cirrhosis	PBC	20	31
Psoriasis	PS	16	35
Primary sclerosing cholangitis	PSCh	6	31
Rheumatoid arthritis	RA	28	81
Systemic Lupus erythematosus	SLE	21	47
Systemic sclerosis	SS	25	6
Type 1 diabetes	T1D	38	57
Ulcerative colitis	UC	45	102

*As reported by Immunobase

Therefore, to prioritize candidate causal genes and variants for IMD I used three different approaches:

1. In chapter three I used an evolutionary approach that allowed us to pinpoint causal SNPs for IMD and gain insight into the pathogenesis of CeD. By intersecting SNPs associated to IMD from the Immunochip with a set of variants inherited from the Neanderthal genome, followed by haplotype analysis, we identified seven IMD loci containing variants inherited from Neanderthal that may affect the risk of developing eight different IMD. We then showed the regulatory potential of the Neanderthal variants within IMD loci. First, we evaluated the effect of the SNPs on nearby genes to define expression quantitative trait loci (eQTL) and pinpointed potential causal genes. Then, we observed that 80.5% of the Neanderthal-variants altered transcription factor binding motifs and showed that the Neanderthal variants in the 14q24.1 locus associated to CeD alter the binding of Rad21, which promotes apoptosis an

important pathway in CeD pathogenesis. Interestingly, CeD was the IMD with the most loci containing Neanderthal variants (3 loci). Finally, we were able to perform fine-mapping of the loci because the regions covering the Neanderthal-haplotype are 10-50% smaller than the locus size.

2. In chapter four we showed that 40% of IMD-SNPs affect the expression of nearby protein-coding genes. However, little was then known about long non-coding RNA genes (lncRNAs) and their role in IMD pathogenesis. Therefore, in chapter four, we evaluated the eQTL effect of IMD using RNA-sequencing (RNA-seq) transcriptomic data from 629 blood samples. RNA-seq data provides quantification of the global transcriptome at high resolution, which allowed us to not only increase the power to identify eQTL, but also to look at transcripts with low expression such as lncRNAs. Using this approach, we prioritized 233 genes from 120 IMD-associated loci, including 53 lncRNAs, which highlights their importance for IMD. As little was known about the function of lncRNAs, we used different layers of genomic information and GeneNetwork[3], a co-expression based network, to predict the function of IMD-associated lncRNAs. By doing so we were able to provide more insights into the functional role of lncRNAs in autoimmunity.

3. In chapter eight, by performing haplotype analysis in four different populations, followed by functional annotation of the candidate variants, we pinpointed one candidate causal variant in the *LPP* (LIM Domain Containing Preferred Translocation Partner In Lipoma) locus, one of the strongest-associated CeD loci. After performing haplotype analysis, we could refine the loci to a 2.8kb region and seven candidate variants. We then annotated these variants using data from the 19 cell lines in the ENCODE project. One of the variants (rs4686484) overlapped with a regulatory region that contains DNase hypersensitive sites (DHSs) and B-cell-specific enhancers and interferes with transcription factor binding sites. This led us to suggest that this SNP might impact gene expression

in B cells. However, in contrast to the data from only 19 cells lines from the ENCODE project, we now have data from 129 different cell lines from the Roadmap epigenomics project in which our prioritized SNP overlaps enhancer histone marks in 39 cell lines, including key cell types in CeD pathogenesis such as B cells, T cells, duodenum mucosa and smooth muscle. We did not find any eQTL effect on *LPP* by SNP rs4686484 or any other SNP in the locus. Only one SNP in moderate LD with rs4686484 had an eQTL effect, which was on *BCL6* (B Cell CLL/Lymphoma 6) in B cells.

After looking at 16 different eQTL studies, we did not find any eQTL in *LPP*, but found that three SNPs in high LD (r2 = 0.8-0.99) with rs4686484 were affecting the expression of *BCL6* in peripheral blood monocytes and one SNP also affected the expression of serine dehydratase like gene (*SDSL*) in the same cell line. Nevertheless, little is known about the function of this gene in the context of CeD. Hence, further studies using disease-specific cell types are needed to understand the function of the prioritized variant and the gene.

Pleiotropy and the importance of looking at disease-specific cell types

In chapter one we observed a large overlap between the loci associated to different IMD. However, our understanding of this pleiotropy was limited because we only knew the associated region, making it unclear whether it was the same genes in these loci affecting different diseases. After our prioritization of candidate causal variants and genes in chapter four, we observed that, although the same physical regions were associated with different diseases, the associated SNPs and affected genes might not be the same. This is because the SNPs associated to different diseases may be located within the same physical region but are not in LD with each other, indicating a role for different haplotypes in different diseases. Moreover, the associated SNPs might also affect different genes, and this might be caused by cell-type-specificity, as shown in our DHSs analysis where SNPs from one disease overlapped DHSs in one cell type, while SNPs from other disease overlapped DHSs in a different cell type. This

also highlights the importance of investigating disease-specific cell types to gain a better understanding of the disease biology. In recent years many publicly available resources containing annotations from different cell types, such as gene expression data, epigenomic marks and DHSs, have become available, and we used these resources for the analysis presented in this thesis. However, this data is still limited to a few cell types and it has been shown that eQTL vary according to cell type and context, such as the state of the cell (activated or not) and different types of stimulations. Given the role of wide range of cell types and conditions in IMD, it is critical to have eQTL information from disease-relevant cell types and conditions.

Secondly, for some diseases we observed that even though they share the same Top-SNP (the most-associated SNP within the LD block), the direction of the effect is opposite, which indicates that the same allele can be protective for one disease while conferring risk to the other. Therefore, the expression of the gene might be increased in one disease, while decreased in the other. We should therefore be careful when interpreting these eQTL results and future studies should always specify the risk allele. Another aspect that has not been explored thus far is the contribution of rare variants to the pleiotropy puzzle, as it has been shown that large effect size susceptibility loci tend to be phenotype-specific[4]. It was hypothesized that rare variants with stronger effect sizes might contribute to the genetics of IMD. However, only a limited number of rare variants have been identified thus far (as partially reviewed in chapter one). To identify these rare variants will require sequencing large numbers of cases and controls. As more sequencing data becomes available, it will be easier to identify these variants and evaluate their true contribution to the disease. This might allow us to gain more insight into the specific mechanisms for each disease beyond just immunity.

Family-based sequencing studies: an alternative method to identify rare variants for complex diseases

Another way to identify rare variants that contribute to disease heritability is by sequencing families with affected individuals in multiple generations. This approach has been successful in many rare diseases, and it has been proposed as an economical way to identify rare variants that explain heritability of complex diseases[5] as some complex diseases, such as CeD, present a dominant-like inheritance in multi-generational families. The proposed sequencing approach for multi-generational families involves initial sequencing of two affected members of the family based on the assumption that both will share the same causal variant. To reduce the number of overlapping candidate variants, the two individuals must be the most distantly related affected family members because more distantly related co-affected individuals will share fewer genetic variants reducing the list of candidate variants. This should be followed by annotation of the candidate variants and examination of their co-segregation with the disease in the rest of the family.

Chapter six describes our efforts to identify rare variants contributing to CeD heritability following this approach. In the first stage of this project we performed whole-exome sequencing (WES) in two individuals from 23 multi-generation families. After functional annotation and filtering of the variants based on MAF and the functional consequence of the variant in coding and regulatory regions, we ended up with a list of 100-120 candidate variants. We then focused on the variants that were present in candidate loci such as case-control associated regions or family-based regions from linkage analysis or homozygosity mapping. Although the best approach would have been to perform exome-wide analysis, this was rather complicated because the resulting list of candidate variants was too large to validate all variants and perform Sanger' sequencing to look at the co-segregation in all members of the family. Additionally, although causal variants should be present in all affected individuals in Mendelian diseases, for complex diseases we cannot exclude variants present in healthy controls because the expected causal variants are not fully penetrant. In appendix I[6], I show the results from our analysis of

one of these multigenerational families. However, this study has a few limitations, the main being the low coverage of our WES data.

In the second stage of this project we performed a better WES in multiple individuals of two large multi-generational families. The large number of affected members in the family permitted us to perform linkage analysis, which reduced d the number of candidate causal variants. By sequencing affected and unaffected individuals (mainly spouses), we were able to look at co-segregation directly and reduce the list of potential causal variants to be followed-up with Sanger sequencing. Results from the analysis of one of the families (CD0605) are presented in the second part of chapter six. After filtering and looking at the co-segregation of the candidate variants, we identified two potential causal variants, one in the *UNC13B* gene and the second in *SPAG8*. Interestingly, we identified another family segregating IgA-deficiency, CeD and common variable immunodeficiency with missense variants in the same two genes, where the variant in *SPAG8* affects the same amino acid that was affected in family CD0605. Unfortunately, we only have DNA from two members of the family, neither of whom had CeD, and it was not possible to re-contact the family to look at the co-segregation of the variants. At this point both genes remain possible causal genes, with further functional studies needed to evaluate their role in the disease. Nevertheless, this approach seems to be better at identifying rare but potentially causal variants for complex diseases.

As WES only covers the exome, and the majority of variants associated to complex diseases are in regulatory regions of the genome, in the third stage of the project we performed whole-genome sequencing (WGS) of 52 individuals who are members of 5 families. Three of the families segregate only CeD, and we aimed to identify rare variants that contribute to the disease. The other two families segregate multiple IMD. Due to the high pleiotropy in IMD, we hypothesized that we would find some shared loci. The major limitation at that time was that most of the non-coding part of the genome was not well annotated, thus we focused our analysis only on the protein-coding part. However, this data will be re-analyzed soon as new data for proper annotation has become available.

This project also demonstrated the importance of keeping in contact with the families and following-up the disease status of the rest of the family members. Many of the DNA samples from our cohort were collected few years before our study was initiated, and some of the patients that we had labeled as healthy might have since developed the disease. The project also highlighted the importance of collecting other types of patient material, such as biopsies in the case of CeD, because these samples can facilitate the functional studies needed to evaluate the role of the candidate variants.

The closest gene is not always the causal gene

With the initial results from the GWAS, it was speculated that one gene was affected per locus and the closest gene to the Top-SNP was reported as a candidate gene. However, in chapter four, by integrating different layers of genomic information, we show that the closest gene is not the causal gene in more than 50% of the loci[7]. Moreover, we show that 39% of SNPs affect multiple genes (multi-SNPs), and that these genes are expressed in different cell types, suggesting that multiple genes in a single locus could contribute to disease through different cell types. We also showed that our multi-SNPs were enriched for super-enhancers and CTCF binding sites that are important for chromatin looping. Additionally, we found that for some loci where the IMD-SNPs are affecting lncRNAs, the affected lncRNAs and the promoters of the protein-coding genes are organized in the same transcription topological unit mediated by RNA Polymerase II. This supports the idea that co-regulation of multiple genes in the locus may occur through looping interactions. We also show an example in which a SNP that affects a lncRNA also affects multiple genes in *trans*, suggesting that this might be another mechanism of action, adding another layer of complexity to the interpretation of results from the association analysis.

Power limitations due to sample size and complications in collecting large cohorts for rare syndromes like TTP

The replication and discovery of new loci via Immunochip has been successful for many of the IMD studied, but the results depend strongly on the prevalence of the disease, the population under study, the cohort size, the relative risk conferred by each of the loci under analysis (the genetic architecture of the disease) and the proximity of the causal variant to the interrogated SNP markers on the Immunochip array[8]. In chapter five, we show how these factors might influence the power to discover new loci in seven diseases with different prevalence and genetic architecture. So far only three diseases out of the 15 studied in this thesis have more than 100 associated loci: inflammatory bowel disease (IBD), multiple sclerosis (MS) and rheumatoid arthritis (RA). Not surprisingly, these are also the three diseases for which we have the largest sample sizes. So, the bigger the sample size, the stronger the study power to discover new and significant associations.

Although attaining a larger sample size is not a limitation for common diseases, where many samples are available, it can be a major complication for rare immunological syndromes such as thrombotic thrombocytopenic purpura (TPP). TTP has a reported incidence between 1 and 13 cases per million people, depending on the geographic location[9]. In chapter two, we performed the first study applying high-throughput DNA-chip genotyping technology to identify genetic risk factors associated with acquired TTP in a cohort of 186 patients and 1,255 controls. Although, this is a large cohort compared to other genetic studies in TTP, the power to detect new associations was limited. While we identified five loci out of the human leukocyte antigen (HLA) region with suggestive P values ranging from 1.59×10^{-5} to 7.6×10^{-5}, none showed a strong association in the independent replication cohort of 88 Italian cases and 456 Italian controls. Nevertheless, we were able to confirm the association to the HLA region (rs6903608, OR = 2.57, P value = 1×10^{-19}). Our results indicate that the HLA class II locus at band 6p21.3 is the main genetic risk factor for acquired TTP, at least among the variants included in the Immunochip. Additionally, the dense genotyping on the Immunochip allowed us to

perform imputation of the classic HLA class I and II genes. Our analysis showed that the Top-SNP in the region (rs6903608) and the HLA allele HLA-DQB1*05:03 explained most of the HLA association with acquired TTP in our discovery population. This finding suggested that rs6903608 probably tags the effect of other HLA alleles, including the widely reported association to HLA-DRB1*11. The independent association with DQB1*05:03 was not observed in previous studies. Further studies with larger samples are needed for not only to confirm the association to DQB1*05:03 but also to identify novel non-HLA loci.

Epistatic effects contribute to 'missing' heritability

Most of the models used to estimate disease heritability are pure genetic models that may underestimate the interactions among loci[10], a phenomenon globally designated as epistasis. The model most often used to estimate the heritability explained by GWAS is the additive model, which simply adds up the effect of the associated variants. However, it can also be that the combination of two or more variants results in a stronger amplified effect that would be missed by the additive model. If this is the case, even if we identify all the genetic variants, we would not be able to explain all the heritability with the additive model, a gap referred to as the 'phantom heritability'. Hence, epistatic components need to be integrated into heritability calculations through estimates of the contribution of non-genetic factors[11]. In IBD, for instance, the estimation of heritability explained by an additive model that includes all 71 loci known from GWAS is 21.5%. However, after considering genetic interactions within the heritability model, the phantom heritability is 62.8%, indicating that genetic interactions could account for 80% of the currently missing heritability in IBD[10].

To test whether some proportion of CeD heritability can be explained by epistatic components, we performed a pilot study to look for interactions between the Top-SNPs in each non-HLA loci using the CeD cohort described in chapter nine. We identify that the locus 2q12.1(rs990171) containing the *IL18R1* and *IL18RAP* genes interacts with two loci: one at 11q23.1 (rs7104791, P = 4.386×10^{-05}) containing *POU2AF1* and

C11orf93 and the other at 16p13.13 (rs9673543, P = $5.526x10^{-05}$) containing *CIITA, SOCS1, PRM1* and *PRM2*. Interestingly, rs990171, the Top-SNP in locus 2q12.1, has an eQTL effect on *IL18RAP, TMEM182, IL18R1* and three ncRNAs (*AC007278.2, AC007278.3* and *MIR4772*). These ncRNAs might be affecting the expression of the protein-coding genes in the other two loci, but further studies are needed to validate this. It is important to perform validation studies using multiple disease relevant cell lines, appropriate stimulations and different time points to be able capture these interactions either at RNA level and/or at protein level. For example, it has been shown that both *IL18RAP* and *POU2AF1* are strongly expressed only after 1 hour of Th2 cell differentiation, and their expression is switched off after 24–48 hours[12]. Our results from this pilot analysis suggest that epistatic effects are present within CeD loci and should be systematically investigated at genome-wide level to reveal more biologically plausible interactions.

New loci change our understanding of disease biology

The identification of new loci associated to CeD, as well as our efforts to pinpoint causal variants and genes, have led to the identification of novel pathways associated to different IMD.

In chapter four, we identified a strong eQTL-effect on the expression of *ULK3*, which encodes a kinase involved in autophagy. Interestingly, using CeD biopsies, we also found an enrichment of autophagy genes being differentially expressed compared to a random set of genes (P = $2.2x10^{-16}$). Furthermore, the *ULK3*-affecting SNP is correlated with the expression levels of autophagy genes in CeD biopsies and that its genotype affects the levels of IL-6 in response to LPS. *ULK3*-dependent autophagy might be involved in regulation of inflammation, which emphasizes the importance of studying non-gluten antigens (e.g. host-microbiome interaction) in the context of CeD pathogenesis[13].

In chapter eight, the discovery of new loci followed by functional annotation of the SNPs and pathway enrichment analysis on the prioritized genes led to the identification of novel CeD pathways including: tumor necrosis

factor mediated signaling, response to tumor necrosis factor, regulation of I-κB kinase/NF-κB signaling, positive regulation of I-κB kinase/NF-κB signaling and apoptotic signaling. Although the NF-κB pathway is a well-known player in CeD pathogenesis, we were the first to show that the dysregulation of this pathway can be one of the causes of CeD rather than just a consequence of the disease.

In chapter three, our evolutionary approach suggested that apoptosis plays a role in CeD, as the many of the Neanderthal-inherited variants overlapped Rad21, which promotes apoptosis. An increased number of apoptotic intestinal epithelial cells in the mucosa of CeD patients has been reported [14–16], but apoptosis had not been implicated by genetics before. However, apoptosis was another novel pathway identified after discovery of new loci in chapter eight, confirming its causal role in CeD pathogenesis.

The discovery of new disease-associated pathways can also change our understanding of disease biology. In CeD, for example, it was initially thought that the immune response was mainly adaptive. However, the Immunochip analysis identified an association to Interleukin 1 Receptor Associated Kinase 1 (*IRAK1*). IRAK1 plays an important role in not only initiating but also regulating innate immune response against pathogens, which suggests that innate immunity also plays a role in CeD. Do other innate immune genes or pathways contribute to CeD? To answer this question, we still need to perform large-scale GWAS as the Immunochip covered less than 4% of the genome.

Discovery of new loci allows the identification of new drug targets

The identification of new loci can also help us identify new drug targets. For instance, a recent trans-ethnic meta-analysis of RA identified 42 new loci and a follow up functional genomics approach led to the identification of 98 candidate causal genes. Interestingly, 27 of these 98 genes overlapped drug-target genes for approved RA drugs[17], demonstrating that GWAS

findings can identify drug-targets. Our own work in chapter eight shows that identifying three novel loci associated to CeD and prioritizing 212 candidate causal genes using a functional genomics approach led to the identification of nineteen genes that overlap known drug-targets. While the repositioning of such drugs to CeD may need further investigation, our results can help prioritize drugs for further studies.

Perspectives

Exploration of the entire genome might add to our understanding of pleiotropy and discover non-immune pathways

The majority of recent analyses that have identified new loci associated to IMD, which included the largest sample sizes, have been performed using the Immunochip. The Immunochip contains 196,524 SNPs across 186 GWAS loci, which still leaves most of the genome unexplored. Although we have already seen a high level of pleiotropy for IMD in the results of GWAS studies (34% of 199 loci were shared by at least two diseases in our analysis from chapter one[2]), it is necessary to perform more genome-wide analysis to correctly estimate the pleiotropy of IMD. It has also been suggested that the pathway enrichment results coming from the analysis of loci discovered by the Immunochip might be biased to immune-related pathways because the loci were partially selected for regions harboring these genes. Thus, as shown in chapter one of this thesis, further genome-wide analysis will reveal more disease specific pathways.

Imputation can help capture most genomic variation and the importance of population-specific reference panels

The number of SNPs present on genotyping chips has increased over the years, however is still impossible to capture most of the variation in the genome using this technology. Although WGS does allow us to capture most of the genome variation, it remains expensive to sequence hundreds of thousands of samples. Nevertheless, the imputation of genotypes using reference panels has greatly facilitated the inference of the genotypes of the markers in high LD with the genotyped SNPs. The imputation of

genotypes using reference panels not only helps increase the genome coverage, it also increases the power to discover new associations and allows better fine-mapping of the associated regions.

It is important to mention that only SNPs present in the reference panel can be imputed, thus it is crucial to have appropriate reference panels. In recent years the number of reference samples has increased, resulting in imputations with high accuracy. For example, the Haplotype reference consortium contains a set of 39,235,157 SNPs from 64,976 human haplotypes and allows imputation of variants with 0.1% MAF[18]. In the coming years, more reference panels will become available as the price of sequencing decreases. The UK biobank, for instance, just finished sequencing 100,000 whole genomes from 85,000 individuals and plans to expand this project to sequence 5 million genomes in the next five years. Furthermore, in 2018 thirteen European countries have come together to share one million genomes for research purposes by 2022.

The technical methods used to perform imputation have also improved considerably. This has made it feasible to impute large cohorts with large reference panels, which will allow us to re-analyze samples that have been already genotyped using GWAS chips or Immunochip. Moreover, SNP-array-based GWAS data and WGS data on large sample sizes improves the power to perform statistical fine-mapping[19]. In a recent study in IBD, for example, imputation of Immunochip genotypes from 33,595 individuals with IBD and 34,257 healthy controls allowed the authors to fine-map 94 loci using a Bayesian approach. This identified 18 associations to a single causal variant with >95% certainty and an additional 27 associations to a single variant with >50% certainty. Interestingly, 13 of the 45 variants were non-synonymous and three of them disrupted the binding of one transcription factor. This study also found a variant in the *IL2RA* gene that had been missed in previous studies because it was absent in HaPmap. It had not been possible to fine-map this region using a GWAS dataset imputed with 1000 genomes project, but the dense genotyping using Immunochip, followed by imputation with 1000genomes project made it possible[20]. Similar studies might be performed in the future with even bigger cohorts and reference panels to gain more insight into disease biology.

Dense genotyping followed by imputation using reference panels has also proven to be useful for identifying population-specific variants that confer risk of disease. In a type 2 diabetes (T2D) study in Mexican and other Latin-American populations, the authors obtained genotypes at 1.38 million SNPs with MAF 0.1% using the Illumina OMNI 2.5 array[21], followed by imputation using the 1000 Genomes Project Phase I, which resulted in 9.2 million SNPs. These genotypes were generated from 3,848 T2D cases and 4,366 controls of Native American and European ancestry. They found a novel genome-wide significant locus associated with T2D that contains the solute carriers *SLC16A11* and *SLC16A13* ($P=3.9x10^{-13}$; OR=1.29). Interestingly, the risk haplotype, containing four amino acid substitutions, is rare in European and African samples, but it is present at 50% frequency in Native American samples and 10% in East Asians. Although this study has a relatively small sample size compared to the previous metabochip analysis for T2D in European populations (34,840 cases and 114,981 controls[22]), it was possible to identify novel loci. The population-specific variant also has a higher OR (1.29) than 64 of the 65 loci coming out the Metabochip analysis, showing that population-specific variants might have stronger effect sizes.

Most of the reference panels that are available or recently announced are from European origin, and there are only a limited number of reference panels from non-European populations. Fortunately, we are seeing an improvement in this respect. The Genome Aggregation Database, for example, contains 125,748 WES and 15,708 WGS from unrelated individuals. This data was generated as part of various disease-specific and population genetic studies. Of these, 8,128 are African/African-American, 17,296 are Latino, 5,040 are Ashkenazi Jewish, 9,197 are East Asian and 15,308 are South Asian. These numbers will likely improve in the future as population-specific reference panels or reference panels with a worldwide population are important. As rare variants are, on average, younger than common variants, they cluster more often based on the geographical distribution and are thus more difficult to impute[23]. An example of how a population-specific reference panel can improve imputation was shown in a recent study of the Anabaptist population[24]

(Amish and Mennonite ancestry). Using WGS data from 265 individuals, the authors identified >12 M high-confidence single nucleotide variants and short indels that were not present in the 1000 genomes project. In addition, 43,000 variants that are usually rare in other populations showed higher allele frequencies in this population. This study showed that combining the Anabaptist reference panel and the 1000 genomes data provided better imputation accuracy than either of the panels alone[24]. Consistent with this, a similar study in Ashkenazi Jews using WES data of 5,685 individuals showed that 34% of the protein-coding alleles were more frequent in this population compared with other reference panels[25]. It also showed that some of these variants might explain the higher prevalence of Crohn's disease in this population.

Another limitation coming from the lack of population-specific reference panels is the absence of correct estimation of allele frequencies in these populations and the correct LD blocks. The difference in allele frequencies of population-specific reference panels compared to publicly available reference panels, such as the 1000 genomes project or the Exome Aggregation Consortium[26] (ExAC), presents a problem for the analysis of high-throughput sequencing data in these populations. In our case, for example, although we sequenced 18 Saharawi families, analysis of this data was not possible because we did not have an adequate reference panel to assess the MAF in the population.

As mentioned above, there are still technical limitations to working with non-European populations. One is that the genotyping chips are usually designed only for European populations, leaving out most of the population-specific variation. The Immunochip, for example, was designed using the variants present in the CEU population of the 1000 genome project phase 1, while the last version of the 1000 genomes project now contains variants from 26 populations (2,504 individuals), including five populations with south Asian ancestry, four with mixed American ancestry, seven with African ancestry, five with East Asian ancestry and five with European ancestry. Thus, the creation of population-specific reference panels will allow us to precisely calculate the LD present in the population in order to create appropriate genotyping chips for these populations.

WGS is better than WES, but still difficult to interpret

Most of the high-throughput sequencing data generated at the beginning of this thesis was WES, which captures almost all the variation affecting proteins but misses most of the variants in regulatory regions. As we have seen from the results of GWAS, most of the disease-associated variants identified so far are located in regulatory regions and alter gene expression. However, with advances in sequencing technology and the fall in sequencing costs, a large amount of WGS data has been generated and the number of full genomes is increasing exponentially. Nevertheless, technical challenges to interpreting the non-coding part of the genome remain. Yet understanding it and establishing its mechanisms of action is crucial not only for interpretation of WGS data, but for interpretation of the results from association studies. As discussed above, exploration of the non-coding part of the genome by WGS will facilitate the identification of rare disease variants in regulatory regions such as enhancer regions or regions promoting chromatin looping interactions. These discoveries might lead to a better understanding of disease pathology and the genetic architecture of IMD.

Family-based sequencing studies using WGS will also reveal the contribution of non-coding variants to disease susceptibility. Although the interpretation of non-coding variants is still difficult because there are millions of them, it is now becoming possible to prioritize non-coding variants using genomic and epigenomic information available from multiple cell lines. For example, a recent study on pancreatic agenesis[27] applied homozygosity mapping in three consanguineous families and identified a locus on chromosome 10 that contained *PTF1A* gene. However, by using Sanger sequencing they could exclude mutations in coding and promoter sequences of *PTF1A* and 24 other genes in the region. In order to identify rare or novel homozygous variants and indels within this region, WGS was performed in two affected members of different families. After filtering out all the variants present in 81 controls, they ended up with a list of 2,868 and 3,188 variants in each individual. Of these variants, 8 and 19 were affecting the protein (missense, nonsense, frameshift or essential splice site) but they failed to co-segregate with the disease or the function

of the genes was not related to pancreas development. The authors then looked for mutations that mapped to active regulatory regions from pancreatic endoderm cells derived from human embryonic stem cell. Interestingly, they found a shared variant within the homozygosity region that was ~25kb downstream of *PTF1A*. Additionally, the authors found that six unrelated patients also shared the same mutation and three more had mutations in the same regulatory region. Finally, using chromatin conformation capture experiments in human pancreatic progenitor cells, they showed that this region is an enhancer and it interacts with the promoter of *PTF1A*, confirming its involvement in the disease.

This study shows that, although it is now easier to prioritize variants using annotations from regulatory regions, the resulting list of candidate variants is huge, which makes methods that allow the identification of candidate loci, such as linkage analysis and homozygosity mapping, essential. The collection of large multi-generational families with multiple affected members thus remains important. Additionally, as the annotation of the non-coding part of the genomes improves and we gain a better understanding of gene-regulation, it will be easier to asses and quantify the effect of non-coding variants, allowing their prioritization. Moreover, with advances in high-throughput technology in few years we could analyze multiple variants at the same time at a low cost.

Sample size is key, but collecting multiple phenotypes is equally important

The amount of genetic data available from genotyping or sequencing will increase enormously in the near future, and this will facilitate the identification of more loci. For example, a recent GWAS of atrial fibrillation included more than one million individuals (60,620 cases and 970,216 control) and resulted in the identification of 111 disease-associated loci, 80 of them novel. The challenge for studying common complex diseases is no longer generating the data, but rather collecting as much phenotypic information as possible in order to link the genetic variants with the phenotypes. The creation of Biobanks has allowed the collection of not only more genotypic data, but also environmental and phenotypic data that

can allow us to perform more specialized genetic analysis. For example, for CeD it will be possible to perform association studies stratified based on HLA haplotype, which will allowing the identification of genetic factors associated with a high risk of developing the disease versus low risk of developing the disease in CeD patients. We can also stratify individuals with high risk based on the age of onset, comorbidities with other IMD or start of gluten intake, leading to discovery of new associated loci that can improve our understanding of disease pathogenesis. It is also important to continue conducting cross-disease meta-analysis because it has proven an efficient way of identifying shared loci and new single-disease associations.

In some cases, biobanks also collect and store patient biomaterial such as biopsies, synovial fluids and stool samples. These stored samples permit the generation of other kinds of omics-data. Some biobanks include the option to re-contact biobank participants for follow-up functional studies. With the collection of stool samples, for instance, it is now possible to look at the gut microbiome of cases compared to controls and identify microbiome features associated to disease. Dysregulation of the gut microbiota has been associated to multiple diseases[28], including IMD such as IBD[29], CeD[30,31], MS[32] and RA[33].

Additionally, the combination of genetic data with longitudinal electronic health records has facilitated the linking of genetic variants with multiple phenotypes at the same time and allows for the creation of better models for predicting the risk of developing the diseases. Better phenotype information will also allow the inclusion of the same patient in multiple studies, reducing the cost of GWAS. As the amount of genetic data continues to increase, these kind of studies will likely increase in the future, including larger samples sizes that will have greater power to detect event smaller effects.

Finally, as we have seen in this thesis, the combination of multiple layers of information, such as genetic and epigenetic data or transcriptomics, allow us to prioritize causal variants and genes, a crucial point for interpreting

GWAS results and translating them into treatments. It is therefore fundamental to continue generating this data in multiple cell types and contexts, such as disease stage or stimulations.

Identification of gene-environment interactions and epigenetic modifications might explain the missing heritability and lead to new ways to treat disease

The collection of environmental data will allow us to look at more complex gene-environment interactions that might help to explain missing heritability. Moreover, understating these interactions might lead to new ways to diagnose and treat the disease. The variability in epigenetic and environmental conditions results in phenotypic variability that in some cases is heritable. Epigenetic changes, such as DNA methylation, can affect gene expression in a heritable manner without actually altering the underlying DNA sequences. An inheritance model that incorporates epigenetic inheritance in addition to genetic effects would help to explain the missing heritabilty[11]. For instance, epigenetic inactivation of some height-associated genes has been shown to be functionally equivalent to Mendelian physical loss of the corresponding alleles[35]. Altered DNA-methylation patterns have also been observed to be transmitted from parents who have been exposed to distinct diets[36–38], stress[39,40], trauma[41–43] or drugs[44–47].

Moreover, some epigenetic modifications have been shown to be caused by exposure to environmental toxins. Exposure to mercury, Bisphenol A, Trichloroethylene and TCDD/AHR ligands, for example, has been shown to alter the immune system[48]. Previous studies in other IMD have also demonstrated that epigenetic effects can contribute to disease susceptibility. For instance, a disease risk allele can differentially alter gene expression depending on its parental origin. This mechanism has already been implicated in IgA deficiency[49], atopic dermatitis[50], MS[51] and type 1 diabetes mellitus[52,53], suggesting that it might play and important role in other IMD as well.

Interestingly, the advantage of epigenetic modifications as treatment targets, compared to genetic changes, is that the epigenetic modifications can be reversed using specific inhibitors. For example, a new study in IBD[54] uses histone deacetylase (HDAC) inhibitors (Givinostat and Vorinostat) to maintain intestinal homeostasis during inflammation. The HDAC inhibitors improved trans-epithelial electrical resistance under inflammatory conditions and also blocked the passage of macromolecules across the epithelial monolayer.

High-throughput functional studies to understand the function of multiple disease-associated variants at the same time

Another challenge lies in determining the causal variants from association studies from those in strong LD, and to identify their underlying mechanism. Annotations of the coding and non-coding part of the genome such as the ENCODE project, the Roadmap Epigenomics and, more recently, the FANTOM5 project have allowed the prioritization of variants, means we can now speculate about the potential role of the variants, such as their location within transcription factor binding sites, histone modifications or DNA methylation sites for a number of transformed and primary cell types. These projects performed most of the annotations using cell lines, however it is now possible to profile cells at single-cell resolution using single-cell sequencing. Taking advantage of this technology, a new project called the Human Cell Atlas (HCA)[55] aims to create a human cell atlas that will not only provide us with transcriptomic data, it will also asses a cell's protein molecules and profile the accessibility of the chromatin in millions of individual cells. In the near future, the HCA will allow us to compare healthy reference cells to diseased ones in relevant tissues as well as prioritize appropriate cell types for functional studies.

In the past few years there have also been some major advances in high-throughput functional studies that now allow the interrogation of multiple variants at the same time. Massively parallel reporter assays (MPRA), for example, can screen thousands of potential functional variants in a single assay to determine effects on gene expression. For instance, using the CRE-seq assay[56], the authors performed high-throughput DNA synthesis

to generate and test more than 1000 genetic variants of a 52-bp rhodopsin promoter. A similar assay[57] was also applied to test more than 27,000 variants of two 87-bp inducible enhancers. Nevertheless, the limitation of the reporter assays is that the regulatory function of a variant will be tested only in the context of plasmid DNA, but not in the context of native genomic DNA in which the variant actually exists. This difference might result in spurious results as it does not take into account the interactions between different genomic and epigenetic components[58].

Until recently, efforts to modify genetic material of an organism or cell have been delayed by a lack of specificity and inefficiency, relying on site-directed mutagenesis or recombination-based methods[59]. However, there has been an enormous improvement in the genome-editing technologies in the past few years. The development of engineered nucleases has resulted in an efficient and highly targeted approach to modifying the genetic architecture of a cell. Three genome-editing techniques have been widely used to date: zinc-finger nucleases (ZFNs), transcription activator-like effector nucleases (TALENs), and clustered regularly interspaced short palindromic repeats (CRISPR) with Cas9 nuclease (CRISPR/Cas9)[59]. Genome editing technologies allow the mutation of an individual SNP from one allele to the other for further comparison in functional studies. In a recent study[60], TALENs were used to confirm the functional role of a SNP previously reported to influence prostate cancer risk. The authors compared edited prostate cancer cell line clones with the three different genotypes and unedited cell lines clones and demonstrated that the risk allele altered *RFX6* expression levels 2-fold. Another study[61] used CRISPR-Cas9 to demonstrate that the risk allele of a T2D variant in the *PPARG2* gene showed increased expression of the transcript in a human pre-adipocyte cell strain. Similar studies have been performed for other diseases and the number will surely continue to grow in coming years.

Similar to MRPA, it is now possible to use high-throughput CRISPR screens. In a recent study[62], for example, the authors developed a CRISPR/Cas9 system for rapid insertion of *TNNT2* gene variants into induced pluripotent stem cell-differentiated cardiomyocytes, with the aim

of generating all the catalogued coding variants in the cardiomyopathy-associated locus *TNNT2* for further functional studies to annotate the function of the variants. Although these authors focused on coding variants, it is also possible to elucidate the role of non-coding variants. Thus, it is believed that high-throughput genome and epigenome editing screens may facilitate the characterization of regulatory function of many potential causal variants located within *cis*-regulatory regions[63], helping us to understand the associations from GWAS studies. In the near future it will be not only possible to interrogate all the variants present in the LD block, but also a combination of them, and this might lead to a catalog that could allow the annotation and prioritization of variants based on these type of studies.

Disease-specific cell types allows us to identify causal genes

Although the discovery of new IMD loci has improved enormously, knowledge of how each locus affects the immune system or specific disease cell types is still lacking. To gain more insight into these topics, it is necessary to identify causal SNPs and genes, and to achieve this it is crucial to use disease-relevant cell types because gene expression has been shown to be cell-type- and context-specific. In chapter five, for example, we showed a locus associated to MS and IBD (11q13.1) where the MS-associated SNP affects two protein-coding genes and a lncRNA, while the IBD-associated SNP affects a different lncRNA. The MS-SNP and its close proxies overlap with DHSs of many immune cells, whereas one of the two proxies of the IBD-SNP rs559928 specifically overlaps with DHSs in Caco-2 cells (intestinal epithelial cells), suggesting that different usage of cell-type-specific enhancers could be one mechanism by which different genes could be affected by different IMD-SNPs in a shared disease locus[7].

However, there is still a lack of information regarding the specific cell types involved in each IMD. One approach to overcoming this limitation is the prioritization of cell types based on overlapping disease-variants with epigenetic and epigenomic features in different cells lines, such as histone marks or DHSs. Another approach is to assess the eQTL effect

for disease-associated variants in multiple cell types or using single-cell sequencing to see where the variants are affecting the expression. Additionally, we can assess the expression of prioritized genes in multiple cell lines. It has been suggested that profiling large cohorts of cases and controls with, for example, cell abundance, signaling response and serum cytokine levels[64], could help us to define immune-phenotypes that can aid to the prioritizations of disease-specific cell types.

Organ-on-a-chip as a model to look at complex interactions

Multiple cell types and tissues might be contributing to the disease pathology, making models that can mimic this physiologically relevant interaction between cell types very important. Animal models have been used to address this problem, but their findings do not always fully represent the human scenario because of fundamental differences in gene-environment interactions. Recently, a new technology has become available that can help to overcome this challenge: organ-on-a-chip. Organ-on-a-chip allows the study of multiple human tissues and cell types in one system. They are microfluidic cell culture systems that recapitulate the structure, function, physiology and pathology of living human organs *in vitro*[65]. The study of organs-on-a-chip derived from the genetic material of patients and controls will help us to gain more insight into disease biology. Moreover, creation of organs with different genotypes will help us understand the effect of specific variants in the whole organ. We could, for example, select individuals at high risk of developing the disease and compare them to individuals with low risk under different environmental conditions and study their consequences. Organs-on-a-chip are the most suitable model so far for performing functional studies because they also contain the epistatic effects in their natural state and allow us to also study gene-environment interactions.

Future Directions

The next 5-10 years should see the development of new ways to diagnosis CeD without the need of intestinal biopsies based not only on the presence of antibodies, but including other kind of biomarkers such as microRNAs, urine gluten peptides or the microbiome. Appropriate genetic

risk scores, including HLA haplotypes and non-HLA risk variants, will also be incorporated into the diagnostic models and will help predict disease prognosis in CeD patients. Moreover, new loci will be discovered using genome-wide technologies, either by genotyping followed by imputation or WGS. The large amount of data that will be generated using multiple cell types from both healthy individuals and CeD patients using single-cell transcriptomics and epigenomics data will lead to the identification of multiple causal genes, including a considerable number of non-coding genes. These advances will also help improve our understanding of how gene regulation works, allowing the interpretation of the effect of those genes. Functional studies, such as organ-on-a-chip studies using cells derived from CeD patients, will lead to improvements in our understanding of disease pathogenesis in multiple contexts. They will include different factors such as genetics, microbiome composition or viral stimulus and aid in understanding of the adaptive and innate immune response to gluten in CeD. Finally, these efforts might lead to the identification and development of potential treatments for CeD patients and new ways to prevent disease development.

References:

1 Mills MC, Rahal C. A scientometric review of genome-wide association studies. doi:10.1038/s42003-018-0261-x.

2 Ricaño-Ponce I, Wijmenga C. Mapping of Immune-Mediated Disease Genes. Annu Rev Genomics Hum Genet 2013; 14: 325–53.

3 Fehrmann RSN, Karjalainen JM, Krajewska M et al. Gene expression analysis identifies global gene dosage sensitivity in cancer. Nat Genet 2015; 47: 115–125.

4 Parkes M, Cortes A, van Heel DA, Brown MA. Genetic insights into common pathways and complex relationships among immune-mediated diseases. Nat Rev Genet 2013; 14: 661–673.

5 Cirulli ET, Goldstein DB. Uncovering the roles of rare variants in common disease through whole-genome sequencing. Nat Rev Genet 2010; 11: 415–25.

6 Szperl A, Ricaño-Ponce I, Li J et al. Exome sequencing in a family segregating for celiac disease. Clin Genet 2011; 80: 138–147.

7 Ricaño-Ponce I, Zhernakova D V., Deelen P et al. Refined mapping of autoimmune disease associated genetic variants with gene expression suggests an important role for non-coding RNAs. J Autoimmun 2016; 68: 62–74.

8 Ricaño-Ponce I, Wijmenga C, Gutierrez-Achury J. Genetics of celiac disease. Best Pract Res Clin Gastroenterol 2015; 29: 399–412.

9 Stanley M, Michalski JM. Thrombotic Thrombocytopenic Purpura (TTP). StatPearls Publishing, 2018http://www.ncbi.nlm.nih.gov/pubmed/28613472 (accessed 16 Jan2019).

10 Zuk O, Hechter E, Sunyaev SR, Lander ES. The mystery of missing heritability: Genetic interactions create phantom heritability. Proc Natl Acad Sci 2012; 109: 1193–1198.

11 Trerotola M, Relli V, Simeone P, Alberti S. Epigenetic inheritance and the missing heritability. 2012. doi:10.1186/s40246-015-0041-3.

12 Kumar V, Gutierrez-Achury J, Kanduri K et al. Systematic annotation of celiac disease loci refines pathological pathways and suggests a genetic explanation for increased interferon-gamma levels. Hum Mol Genet 2015; 24: 397–409.

13 Ricaño-Ponce I, Zhernakova D V., Deelen P et al. Refined mapping of autoimmune disease associated genetic variants with gene expression suggests an important role for non-coding RNAs. J Autoimmun 2016; 68: 62–74.

14 Mazzarella G, Stefanile R, Camarca A et al. Gliadin Activates HLA Class I-Restricted CD8+ T Cells in Celiac Disease Intestinal Mucosa and Induces the Enterocyte Apoptosis. Gastroenterology 2008; 134: 1017–1027.

15 Maiuri L, Ciacci C, Raia V et al. FAS engagement drives apoptosis of enterocytes of coeliac patients. Gut 2001; 48: 418–24.

16 Shalimar D, Das P, Sreenivas V, Gupta SD, Panda SK, Makharia GK. Mechanism of Villous Atrophy in Celiac Disease: Role of Apoptosis and Epithelial Regeneration. Arch Pathol Lab Med 2013; 137: 1262–1269.

17 Okada Y, Wu D, Trynka G et al. Genetics of rheumatoid arthritis contributes to biology and drug discovery. Nature 2014; 506: 376–381.

18 McCarthy S, Das S, Kretzschmar W et al. A reference panel of 64,976 haplotypes for genotype imputation. Nat Genet 2016; 48: 1279–1283.

19 Visscher PM, Wray NR, Zhang Q et al. 10 Years of GWAS Discovery: Biology, Function, and Translation. Am J Hum Genet 2017; 101: 5–22.

20 Huang H, Fang M, Jostins L et al. Fine-mapping inflammatory bowel disease loci to single-variant resolution. Nature 2017; 547: 173–178.

21 SIGMAType T, Consortium D. Sequence variants in SLC16A11 are a common risk factor for type 2 diabetes in Mexico. Nature 2014; 506. doi:10.1038/nature12828.

22 Morris AP, Voight BF, Teslovich TM et al. Large-scale association analysis provides insights into the genetic architecture and pathophysiology of type 2 diabetes. Nat Genet 2012; 44: 981–90.

23 Bomba L, Walter K, Soranzo N. The impact of rare and low-frequency genetic variants in common disease. doi:10.1186/s13059-017-1212-4.

24 Hou L, Kember RL, Roach JC et al. A population-specific reference panel empowers genetic studies of Anabaptist populations. Sci Rep 2017; 7: 1–9.

25 Kenny EE, Pe'er I, Karban A et al. A genome-wide scan of ashkenazi jewish crohn's disease suggests novel susceptibility loci. PLoS Genet 2012; 8. doi:10.1371/journal.pgen.1002559.

26 Lek M, Karczewski KJ, Minikel E V. et al. Analysis of protein-coding genetic variation in 60,706 humans. Nature 2016; 536: 285–291.

27 Weedon MN, Cebola I, Patch A-M et al. Recessive mutations in a distal PTF1A enhancer cause isolated pancreatic agenesis Europe PMC Funders Group. Nat Genet 2014; 46: 61–64.

28 Lynch S V., Pedersen O. The Human Intestinal Microbiome in Health and Disease. N Engl J Med 2016; 375: 2369–2379.

29 Hold GL, Smith M, Grange C, Watt ER, El-Omar EM, Mukhopadhya I. Role of the gut microbiota in inflammatory bowel disease pathogenesis: what have we learnt in the past 10 years? World J Gastroenterol 2014; 20: 1192–210.

30 Marasco G, Di Biase AR, Schiumerini R et al. Gut Microbiota and Celiac Disease. Dig Dis Sci 2016; 61: 1461–1472.

31 Wacklin P, Kaukinen K, Tuovinen E et al. The duodenal microbiota composition of adult celiac disease patients is associated with the clinical manifestation of the disease. Inflamm Bowel Dis 2013; 19: 934–41.

32 Calvo-Barreiro L, Eixarch H, Montalban X, Espejo C. Combined therapies to treat complex diseases: The role of the gut microbiota in multiple sclerosis. Autoimmun Rev 2018; 17: 165–174.

33 Horta-Baas G, Romero-Figueroa M del S, Montiel-Jarquín AJ, Pizano-Zárate ML, García-Mena J, Ramírez-Durán N. Intestinal Dysbiosis and Rheumatoid Arthritis: A Link between Gut Microbiota and the Pathogenesis of Rheumatoid Arthritis. J Immunol Res 2017; 2017: 1–13.

34 Furrow RE, Christiansen FB, Feldman MW. Environment-sensitive epigenetics and the heritability of complex diseases. Genetics 2011; 189: 1377–1387.

35 Tripaldi R, Stuppia L, Alberti S. Human height genes and cancer. Biochim Biophys Acta 2013; 1836: 27–41.

36 Dolinoy DC, Huang D, Jirtle RL. Maternal nutrient supplementation counteracts bisphenol A-induced DNA hypomethylation in early development. Proc Natl Acad Sci U S A 2007; 104: 13056–61.

37 Waterland RA, Kellermayer R, Laritsky E et al. Season of Conception in Rural Gambia Affects DNA Methylation at Putative Human Metastable Epialleles. PLoS Genet 2010; 6: e1001252.

38 Painter RC, Osmond C, Gluckman P, Hanson M, Phillips DIW, Roseboom TJ. Transgenerational effects of prenatal exposure to the Dutch famine on neonatal adiposity and health in later life. BJOG 2008; 115: 1243–9.

39 Babenko O, Golubov A, Ilnytskyy Y, Kovalchuk I, Metz GA. Genomic and epigenomic responses to chronic stress involve miRNA-mediated programming. PLoS One 2012; 7. doi:10.1371/journal.pone.0029441.

40 Zannas AS, Arloth J, Carrillo-Roa T et al. Lifetime stress accelerates epigenetic aging in an urban, African American cohort: relevance of glucocorticoid signaling. Genome Biol 2015; 16: 266.

41 Perroud N, Rutembesa E, Paoloni-Giacobino A et al. The Tutsi genocide and transgenerational transmission of maternal stress: epigenetics and biology of the HPA axis. World J Biol Psychiatry 2014; 15: 334–345.

42 Dickson DA, Paulus JK, Mensah V et al. Reduced levels of miRNAs 449 and 34 in sperm of mice and men exposed to early life stress. Transl Psychiatry 2018; 8: 101.

43 Youssef N, Lockwood L, Su S et al. The Effects of Trauma, with or without PTSD, on the Transgenerational DNA Methylation Alterations in Human Offsprings. Brain Sci 2018; 8: 83.

44 Golding J, Ellis G, Gregory S et al. Grand-maternal smoking in pregnancy and grandchild's autistic traits and diagnosed autism. Sci Rep 2017; 7: 46179.

45 Accordini S, Calciano L, Johannessen A et al. A three-generation study on the association of tobacco smoking with asthma. Int J Epidemiol 2018; 47: 1106–1117.

46 Chastain LG, Sarkar DK. Alcohol effects on the epigenome in the germline: Role in the inheritance of alcohol-related pathology. Alcohol 2017; 60: 53–66.

47 Pembrey ME, Bygren LO, Kaati G et al. Sex-specific, male-line transgenerational responses in humans. Eur J Hum Genet 2006; 14: 159–166.

48 Blossom SJ, Gilbert KM. Epigenetic underpinnings of developmental immunotoxicity and autoimmune disease. Curr Opin Toxicol 2018; 10: 23–30.

49 Vorechovský I, Webster AD, Plebani A, Hammarström L. Genetic linkage of IgA deficiency to the major histocompatibility complex: evidence for allele segregation distortion, parent-of-origin penetrance differences, and the role of anti-IgA antibodies in disease predisposition. Am J Hum Genet 1999; 64: 1096–109.

50 Esparza-Gordillo J, Matanovic A, Marenholz I et al. Maternal Filaggrin Mutations Increase the Risk of Atopic Dermatitis in Children: An Effect Independent of Mutation Inheritance. PLoS Genet 2015; 11: 1–16.

51 Ramagopalan S V., Yee IM, Dyment DA et al. Parent-of-origin effect in multiple sclerosis. Neurology 2009; 73: 602–605.

52 Bennett ST, Wilson AJ, Esposito L et al. Insulin VNTR allele-specific effect in type 1 diabetes depends on identity of untransmitted paternal allele. Nat Genet 1997; 17: 350–352.

53 Wallace C, Smyth DJ, Maisuria-Armer M, Walker NM, Todd JA, Clayton DG. The imprinted DLK1-MEG3 gene region on chromosome 14q32.2 alters susceptibility to type 1 diabetes. Nat Genet 2010; 42: 68–71.

54 Friedrich M, Gerbeth L, Gerling M et al. HDAC inhibitors promote intestinal epithelial regeneration via autocrine TGF 1 signalling in inflammation. Mucosal Immunol 2019; : 1.

55 Rozenblatt-Rosen O, Stubbington MJT, Regev A, Teichmann SA. The Human Cell Atlas: from vision to reality. Nature 2017; 550: 451–453.

56 Kwasnieski JC, Mogno I, Myers CA, Corbo JC, Cohen BA. Complex effects of nucleotide variants in a mammalian cis-regulatory element. Proc Natl Acad Sci U S A 2012; 109: 19498–503.

57 Melnikov A, Murugan A, Zhang X et al. Systematic dissection and optimization of inducible enhancers in human cells using a massively parallel reporter assay. Nat Biotechnol 2012; 30: 271–277.

58 Inoue F, Ahituv N. Decoding enhancers using massively parallel reporter assays. Genomics 2015; 106: 159–164.

59 Smith AJP, Deloukas P, Munroe

PB. Emerging applications of genome-editing technology to examine functionality of GWAS-associated variants for complex traits. Physiol Genomics 2018; 1: physiolgenomics.00028.2018.

60 Spisák S, Lawrenson K, Fu Y et al. CAUSEL: an epigenome- and genome-editing pipeline for establishing function of noncoding GWAS variants. Nat Med 2015; 21: 1357–1363.

61 Claussnitzer M, Dankel SN, Klocke B et al. Leveraging cross-species transcription factor binding site patterns: from diabetes risk loci to disease mechanisms. Cell 2014; 156: 343–58.

62 Lv W, Qiao L, Petrenko N et al. Functional Annotation of TNNT2 Variants of Uncertain Significance With Genome-Edited Cardiomyocytes. Circulation 2018; 138: 2852–2854.

63 Gallagher MD, Chen-Plotkin AS. The Post-GWAS Era: From Association to Function. 2018. doi:10.1016/j.ajhg.2018.04.002.

64 Gutierrez-Arcelus M, Rich SS, Raychaudhuri S. Autoimmune diseases - connecting risk alleles with molecular traits of the immune system. Nat Rev Genet 2016; 17: 160–74.

65 Bein A, Shin W, Jalili-Firoozinezhad S et al. Microfluidic Organ-on-a-Chip Models of Human Intestine. Cmgh 2018; 5: 659–668.

AM Szperl[a*], I Ricaño-Ponce[a*], JK Lib[b*], P Deelen[a], A Kanterakis[a], V Plagnolc[c], F van Dijk[a], HJWestra[a], G Trynka[a], CJMulder[d], M Swertz[a], C Wijmengaa[a*] and H Ch Zheng[b*]

a. Department of Genetics, University Medical Center Groningen, University of Groningen, Groningen, The Netherlands,

b. Department of Biomedical Research, Research & Cooperation Division, BGI-Shenzhen, Shenzhen, China, cUCL Genetics Institute, University College London, London, UK, and

d. Department of Gastroenterology, VU Medical Center, Amsterdam, The Netherlands

*These authors contributed equally.