Studying complex traits in the post-GWAS era: application to asthma and allergy-related traits

Andréanne Morin

Department of Human Genetics
Faculty of Medicine
McGill University, Montreal, Canada
August 2017

A thesis submitted to McGill University in partial fulfillment of the requirements of the degree of Doctor of Philosophy

Table of Content

Abstract/Résumé

Abstract

In the past few years, genome-wide association studies (GWAS) allowed to identify a large number of common variants associated with multiple complex traits. These studies were a great tool that really helped understanding the genetic basis of a large number of diseases allowing to identify new pathways, better understand disease mechanisms and even pinpoint potential drug targets. However, most of the SNP identified were located in the non-coding region of the genome (~90%) and mainly had small effect size. Additionally, even the largest meta-analysis combining thousands of samples could not explain most of the diseases heritability. This forced the research community to develop new strategies and tools to complement GWAS findings. In this thesis, we explored some of these strategies to study asthma and allergy-related traits. These diseases are highly heterogeneous, having important genetic and environmental components. They affect millions of people around the world resulting in many deaths and consist an important economic burden. We used two strategies to understand the genetic basis of these diseases: 1) exploring the impact of rare and low-frequency variants and 2) using DNA methylation data to understand the functional impact of SNPs. We first developed a custom capture panel to assess both coding and non-coding rare and low-frequency regulatory variants to explore their impact on autoimmune and inflammatory complex traits. We applied it to a familial asthma cohort from a founder population and identified three novel genes associated with related traits (serum IgE levels and eosinophil percentage). We also used DNA methylation data to complement our findings as well as to identify new genes associated with allergic rhinitis. The results presented in this thesis represent a good example on how to learn from GWAS findings and go beyond them to understand the genetic basis of complex traits.

Résumé

Au cours des dernières années, les études d'association pangénomique (GWAS) ont permis d'identifier un grand nombre de variants communs associés à de multiples traits complexes. Ces études ont été un excellent outil et ont largement contribué à comprendre la génétique d'un grand nombre de maladies permettant d'identifier de nouvelles voies biologiques, de mieux comprendre les mécanismes de la maladie et même de cerner de potentielles cibles de médicaments. Cependant, la plupart des SNP identifiés étaient situés dans la région non codante du génome (~ 90%) et avaient surtout un effet limité. En outre, même la plus grande méta-analyse combinant des milliers d'échantillons n'a pas pu expliquer la plupart de l'héritabilité des maladies. Cela a obligé le milieu de la recherche à développer de nouvelles stratégies et outils pour complémenter les résultats des GWAS. Dans cette thèse, nous avons exploré certaines de ces stratégies pour étudier l'asthme et l'allergie. Ces maladies sont très hétérogènes, ayant des composantes génétiques et environnementales importantes. Elles affectent des millions de personnes dans le monde entraînant de nombreux décès et constituent un fardeau économique important. Dans cette thèse, nous avons utilisé deux stratégies pour comprendre la génétique de ces maladies : 1) l'exploration de l'impact des variants rares et de faible fréquence et 2) l'utilisation de la méthylation de l'ADN pour comprendre l'impact fonctionnel des variants. Nous avons d'abord développé une capture personnalisée pour évaluer à la fois les variants régulateurs codants et non-codants, rare et de faibles fréquences, pour explorer leur impact sur les traits complexes auto-immuns et inflammatoires. Nous l'avons appliquée à une cohorte d'asthme familial provenant d'une population fondatrice et avons identifié trois nouveaux gènes associés à des traits apparentés (niveau d'IgE et pourcentage d'éosinophiles). Nous avons également utilisé des données de méthylation de l'ADN pour complémenter nos résultats ainsi que pour identifier de

nouveaux gènes associés à la rhinite allergique. Les résultats présentés dans cette thèse représentent un bon exemple sur la façon d'apprendre des résultats des analyses GWAS et de les complémenter pour comprendre la génétique des traits complexes.

Acknowledgements

I would like to start by thanking my two supervisors Tomi Pastinen and Catherine Laprise. Thank you Tomi for this great opportunity to work on this project. Thank you for giving me a chance even though I had very small experience in bioinformatics. I am grateful for your trust and patience. Thank you also to Catherine who has been my supervisor since my master's degree. Thank you for all of the opportunities and for sharing your passion for research with me.

I would also like to thank my committee members Drs Brent Richards and Simon Gravel for incredible enthusiasm, support and advices throughout my PhD. Thanks to Simon Gravel for his great advice on population genetics and Brent Richards on rare variants analyses.

It has also been a great privilege to work with incredible graduate students, Postdocs and research associates. I would like to thank: Tony Kwan, Bing Ge, Warren Cheung, Albena Pramatarova, Stephan Busche, Véronique Adoue, Adriana Redensek, Nick Light, Fiona Allum, Toby Hocking, XioaJian Shao. Thank you all for your help and advices.

I had a chance to meet amazing people in the Department of Human Genetics. I would like to thank all the students I worked with in the Human Genetics Student Society. I had a great experience learning different leadership and organization skills with you. A special thanks to Renata and Karine for being my sport and stress relief partners. I would also like to thank the department, Aimee Ryan, Ross MacKay, Rimi and Eric Shoubridge. This department really cares about the well-being of its student and I always felt supported. Thank you all for your incredible help.

The completion of this thesis would not have been possible without the constant support of my family. I would like to thank Julien who has been there for me on a daily basis. Thank you for your encouragements, keeping me positive and a better version of myself. I would also like to thank my parents Charles and Marie-Dominique as well as my sister Elisabeth for their constant encouragement and support. I always felt loved, listened and supported since day one. Thanks to my father and sister for great science/research discussion. I hope that we will work together in the future on game changing projects. Thanks to my mom for being my main confident.

Finally, I would like to thank the Canadian Institute of Health Research that funded this work. Throughout my PhD, I was supported by generous Doctoral training award from the Réseau de Médecine Génétique Appliqué (RMGA) and the Fond de Recherche du Québec en Santé (FRQS). Thank you also to the Human Genetic Department for the travel and excellence awards.

List of Abbreviations

1KG: 1000 Genome Project

ACTL9: actin like 9

ADRB2: adrenoceptor beta 2

AGER: advanced glycosylation end-product specific receptor

AI: Allelic imbalance

ANOVA: analysis of variance

AR: allergic rhinitis

ARA: allergic rhinitis with asthma

AS: Allele-specific

ASE: Allele-specific Expression

ASH: Allele-specific histone deposition

ASM: Allele-specific Methylation

ATG5: autophagy related 5

C11orf30: EMSY, BRCA2 interacting transcriptional repressor

CADD: Combined annotation dependent depletion

CCDC126: coiled-coil domain containing 126

CDC123: cell division cycle 123

CDX1: Caudal Type Homeobox 1

CLK2P: CDC like kinase 2, pseudogene 1

CMC: Combined Multivariate and Collapsing

COPD: chronic obstructive pulmonary disease

CRKRS: cyclin dependent kinase 12

CX3CR1: C-X3-C motif chemokine receptor 1

CXCR6: C-X-C motif chemokine receptor 6

DHS: DNase I hypersensitive sites

DNase I: deoxyribonuclease I

DNase-seq: DNase I sequencing

EMMAX: Efficient Mixed Model Association eXpedited

Eos: eosinophil

EPACTS: Efficient and Parallelizable Association Container Toolbox

eQTL: expression Quantitative trait loci

eSNP: SNPs altering expression

EWAS: epigenome-wide association study

FCER1A: Fc fragment of IgE receptor 1a

FDR: false discovery rate

FEV_1: Forced expiratory volume in one second

FIMO: Finding Individual Motif Occurrence

FLG: filaggrin

FVC: Forced Vital Capacity

FYCO1: FYVE and coiled-coil domain containing 1

GATA2: GATA binding protein 2

GERP: Genomic Evolutionary Rate Profiling

GSDMA: gasdermin A

GSDML/GSDMB: gasdermin B

GSTCD: glutathione S-transferase C-terminal domain containing

GRASP: general receptor for phosphoinositides 1 associated scaffold protein

GRCh37: Genome Reference Consortium Human genome build 37

GWAS: Genome-wide Association study

HCG22: HLA Complex Group 22

HGP: Human Genome Project

HHIP: hedgehog interacting protein

HLA: Human Leucocyte Antigen

HLA-C: Major Histocompatibility Complex, Class I, C

HLA-DQA1: Major Histocompatibility Complex, Class II, DQ alpha 1

HLA-DQB1: Major Histocompatibility Complex, Class II, DQ beta 1

HLA-F: Major Histocompatibility Complex, Class I, F

HTR4: 5-hydroxytryptamin receptor 4

IgE: immunoglobulin E

IL1R2: interleukin 1 receptor type 2

IL1RL1: interleukin 1 receptor like 1

IL13: interleukin 13

IL18R1: interleukin 18 receptor 1

IL18RAP: interleukin 18 receptor accessory protein

IL33: interleukin 33

IFNGR1: interferon gamma receptor 1

Kb: Kilo base

KIF3A: kinesin family member 3A

LDL: lymphoblastoid cell line

LD: Linkage Disequilibrium

LoF: Loss of Function

LPP: LIM domain containing preferred translocation partner in lipoma

LRRC32: leucine rich repeat containing 32

MAF: Minor Allele Frequency

Mb: mega base

MHC: Major Histocompatibility Complex

mQTL: methylation quantitative trait loci

MRPL44: mitochondrial ribosomal protein L44

MTHFR: methylenetetrahydrofolate reductase

MUC22: Mucin 22

NGS: Next-generation Sequencing

NOS1: nitric oxide synthase 1

NRP2: neuropilin 2

ORMDL3: ORMDL sphingolipid biosynthesis regulator 3

OVOL1: ovo like transcriptional repressor 1

PCA: principal component analysis

PLAU: plasminogen activator urokinase

PRR5L: proline rich 5 like

RAD50: RAD50 double strand break repair protein

RNA-seq: RNA sequencing

RNF39: Ring Finger Protein 39

SERPINE2: serpin family E member 2

SHMT1: serine hydroxymethyltransferase 1

SKAT: Sequence kernel association test

SLSJ: Saguenay–Lac-Saint-Jean

SMCR8: Smith-Magenis syndrome chromosome region, candidate 8

SNP: Single Nucleotide polymorphism

SNV: Single Nucleotide variation

SWAN: subset-quantile within array normalization

TAGC: Translational Asthma Genetic Consortium

THSD4: thrombospondin type 1 domain containing 4

TLR1: toll like receptor 1

TLR6: toll like receptor 6

TNFRSF6B: TNF receptor superfamily member 6

TNIP1: TNFAIP3 Interacting Protein 1

TSS: Transcription Start Site

Ts/Tv: transition to transversion ratio

UK: United Kingdom

UTR: Untranslated region

WDR36: WD repeat domain 36

WES: Whole-exome Sequencing

WGS: Whole-genome Sequencing

ZFP57: Zinc Finger Protein 57

ZPBP2: zona pellucida binding protein 2

List of Figures

List of Tables

Preface

Contribution of authors

The work described here was performed under the co-supervision of Drs. Tomi Pastinen and Catherine Laprise. It is a manuscript based format thesis as described in the Thesis Preparation Guidelines by the Department of Graduate and Postdoctoral Studies. This thesis contains five chapters. The first chapter is a review of the literature relevant to this thesis. Chapters 2 and 4 are works that have been published in *BMC Medical Genomics* and *Clinical Epigenetics* journals respectively. Chapter 3 is a manuscript that will be submitted to *European Journal of Human Genetics*. Chapter 5 is a general discussion of the findings and future directions followed by an overall conclusion. Finally, a summary of contribution to other projects can be found in the annex section.

Chapter 2 is a manuscript authored by *Andréanne Morin, Tony Kwan, Bing Ge, Louis Letourneau, Maria Ban, Karolina Tandre, Maxime Caron, Johanna K. Sandling, Jonas Carlsson, Guillaume Bourque, Catherine Laprise, Alexandre Montpetit, Ann-Christine Syvanen, Lars Ronnblom, Stephen J. Sawcer, Mark G. Lathrop and Tomi Pastinen.* It was published in *BMC Medical Genomics* in September 2016. TP, GB and ML conceived and supervised the study. AM, TP and TK drafted the manuscript. AM, TK, BG, MC and LL analyzed the data. MB, KT, JS, JC, A-CS, LR, SJS provided samples and materials. All authors reviewed and approved the final manuscript.

Chapter 3 is a manuscript authored by *Andréanne Morin**,** Tony Kwan, Anne-Marie Madore, Maria Ban, Jukka Partanen, Lars Rönnblom, Ann-Christine Syvänen, Stephen Sawcer, Hendrik Stunnenberg, Mark Lathrop, Tomi Pastinen, Catherine Laprise* and will be submitted to the

European Journal of Human Genetics. TP, CL and ML conceived and supervised the study. AM drafted the manuscript. CL collected the data and managed the SLSJ cohort. AM and TK analysed the data. AMM, MB, JP, A-CS, LR, SJS and HS provided samples and materials. All authors reviewed and approved the final manuscript.

Chapter 4 is a manuscript authored by *Andréanne Morin, Michel Laviolette, Tomi Pastinen, Louis-Philippe Boulet and Catherine Laprise*. It was published in January 2017 in *Clinical Epigenetics*. CL collected the data and managed the SLSJ cohort, conceived and supervised the study. AM analyzed and interpreted the data and wrote the manuscript draft under the supervision of CL. CL, LPB, ML, and TP edited the manuscript. All authors reviewed and approved the final manuscript.

Original contribution to knowledge

This thesis explores different strategies to better understand genetics underlying complex traits and more specifically asthma and allergy related traits. We used both rare and low-frequency variants exploration as well as understanding the impact of genetics on cellular traits (in this case, gene expression and DNA methylation).

The first study described in Chapter 2 is entitled "Immune-genetics sequencing: the identification of functionally relevant variants through targeted capture and sequencing of active regulatory region in human immune cells". It describes how we built a custom capture panel (Immune-genetics sequencing) that targets regulatory regions of immune cells in order to study rare variants in autoimmune and inflammatory complex traits. We assessed the functional impact of variants identified using our custom capture panel in 30 healthy samples for which we also had gene expression data (RNA-sequencing) in T-cells. We took advantage of this data to evaluate the impact of rare and novel variants on gene expression. We showed that our Immune-genetics sequencing was properly designed and that it identified rare and novel variants that have a high potentially functional impact and that have an influence on gene expression in T-cells. We finally replicated our data in a sample set of 180 individuals.

The second study described in Chapter 3 is entitled "Exploring rare and low-frequency variants in the Saguenay–Lac-Saint-Jean population identified genes associated with asthma and allergy related traits". We used our Immune-genetics sequencing described in Chapter 2 on 149 trios from the SLSJ asthma familial cohort. We first assessed the characteristic of the rare and low-frequency variants in this founder population compared to four other European population from

France, United Kingdom (UK), Sweden and Finland; the latter also being a founder population. We observe a small enrichment of deleterious variants in the low-frequency spectrum in the two founder populations. We were also able to observe a higher proportion of private variants that reached testable frequencies in the SLSJ and Finland populations. We next looked at the impact of rare variants on asthma and allergy related traits in the SLSJ asthma familial cohort testing for lung function, Immunoglobulin (Ig) E levels and eosinophil percentage. Using single variants association test, we identified a low-frequency variant located between *CXCR6* and *FYCO1* genes significantly associated with eosinophil percentage. We also used gene-based test where we identified two genes significantly associated with eosinophil percentage (*MRPL44)* and serum IgE levels (*NRP2*). None of the genes we discovered were previously associated with the traits.

Finally, the third study described in Chapter 4 is entitled "Combining omics data to identify genes associated with allergic rhinitis". This paper on the SLSJ asthma familial cohort focused on allergic rhinitis trait, an asthma related phenotype. We combined Genome and Epigenome Wide Association studies (GWAS and EWAS) using methylation quantitative trait loci (mQTLs) to identify new genes associated with the trait. We were able to identify the *CDX1* gene that was not associated with the trait in prior study.

Chapter 1: Introduction

1.1 Complex traits: asthma and allergic diseases as an example

Complex traits are common diseases that tend to cluster in families and have a large genetic component, thus being heritable. However, they are also greatly influenced by environmental factors, which make them highly heterogeneous. Asthma is an example of complex traits that affect millions of people worldwide. It comprises multiple subphenotypes including allergic asthma. The latter is part of a process called the "atopic march", which is a process starting from early life allergic sensitization leading to the development of asthma, allergic rhinitis or both later in life. The genetic components of these traits have been highly studied allowing the discovery of hundreds of genes associated with them. In this thesis, I will first describe how the genetic background of complex traits has been studied, first by focusing mainly on genome-wide association studies (GWAS) that were broadly popular and utilized in the past. I will then focus on what was learned from them and new strategies to understand the genetic aspects that GWAS could not uncover. I will then finish by describing asthma and allergy-related traits and what is known so far about their genetic components.

1.2 The Genome-Wide Association Study (GWAS) era

GWAS became popular in the mid-2000s and were an important step forward in the study of complex traits. These traits were previously studied using either linkage analyses or candidate gene studies. The former was better suited to identify loci of Mendelian or monogenic diseases, and was not as successful in deciphering the genetic contribution to complex traits [1]. It relied on the inheritance pattern in families and the unclear pattern of Mendelian inheritance that

complex traits harbored made them difficult to study using this technique [1]. Other caveats such as the low power (for variants of small effect size) and low resolution pointing to large regions of the genome, made results hard to reproduce [1]. In the case of candidate gene studies, the main obstacle resided in the limited number of variants selected based on *a priori* knowledge of biological pathways linked to the pathophysiology [2, 3]. The findings were limited to what was previously known but were still hard to replicate [2, 3].

The advent of lower-cost genotyping chips led to GWAS supplanting linkage and candidate gene studies, and allowing for cost-effective interrogation of hundreds of thousands of markers across the genome in larger cohorts. GWAS is a method to test the association of multiple single-nucleotide polymorphisms (SNPs) with a trait simultaneously. It was the first way to explore the whole genome in a cost-effective manner. It started with the completion of the Human Genome Project (HGP) in 2000, which resulted in the first draft of the human genome [4, 5]. This map gave researchers a reference sequence and served as a great starting point for the discoveries that followed. One of the main uses of the HGP came during the GWAS era. Around that time, progress in microarray technology design came to a point where thousands of common variants could be assessed simultaneously in a large number of samples. The design of the genotyping chip was helped by the International HapMap project, which in 2003 identified the majority of common SNPs interrogated in GWAS [6]. Using these arrays in combination with imputation from a reference panel (either from HapMap or more recently from 1000 Genomes project) allowed identifying a large number of genetic loci predisposing for complex diseases. The first successful GWAS was published in 2005 on Age-related macular degeneration [7] and was followed by thousands of GWAS studies assessing mostly common complex traits [8]. These efforts identified hundreds of genetic variants associated with different traits and have been

reported in the GWAS catalog (NHGRI) [8]. Overall, GWAS provided good insight into the complex genetic architecture of these traits [9]. Some of them mainly differ on the number of causal variants as well as their frequency and effect size [9]. They also played an important role in the discovery of novel biological pathways leading to a better understanding of causal mechanisms of diseases [10]. They also identified key elements for disease prediction [11, 12] and helped uncover new potential drug targets for a plethora of diseases [13].

Even though GWAS represented a great step forward in understanding the genetic contribution to a large number of diseases, they did not lead to the identification of high effect size variants allowing for disease prediction. Association studies and subsequent meta-analysis studies reaching hundreds of thousands of subjects to increase the statistical power lead to the identification of a large number of variants, but most of them had small effect sizes and, even when combining them, did not explain a large part of the heritability [14]. Another downside is that the majority of the identified SNPs are located in the non-coding region of the genome, both intronic and intergenic regions (>90%), making it hard to pinpoint the relevant gene [15]. GWAS also mainly assess common SNPs, leaving variants with lower frequencies unexplored. In addition, GWAS SNPs lie in large haplotype blocks of multiple SNPs in linkage disequilibrium meaning that these variants are most often transmitted concomitantly and equally associated to the traits, making it hard to identify the causal variants and underlying biological mechanism. Finally, the stringent genome-wide significance threshold to limit false discovery due to the high number of tests performed could also lead to false negative findings where many true associations failed to reach the threshold and thus are never being considered for further investigation [16]. To complement GWAS caveats in an effort to explain part of the "missing heritability" and to get better insight into disease pathophysiology, studies have started using

different strategies: assessing 1) functional impact of variants [17, 18], 2) rare and low-frequency variations [9], 3) epistasis (gene-gene interaction, [19]) 4) gene-environment interactions [20] and 5) structural variations [21]. I focus on the first two strategies in the body of this thesis.

1.2.1 "Missing" or "hidden" heritability in complex traits

Heritability is a concept that can be summarized as the estimation of the degree to which genetic factors explains a phenotype [22] and is usually determined through twin or family studies. Following the advent of GWAS came a large focus on the so-called "missing heritability" of diseases. All together, significant variants identified through GWAS only explain a small fraction of the genetic variance with the remaining unexplained heritability referred to in past years as the "missing heritability" [14]. Others also suggested that the so-called "missing heritability" was more of a "hidden heritability" because it was not detected due to the stringent multiple testing corrections used in GWAS [23]. Studies on height have shown that they could explain more of the trait heritability when taking all common SNPs into account instead of focusing on the significantly associated ones [24]. Common variants residing out of the reach of GWAS studies are also thought to contribute to this missing heritability along with rare and low-frequency variants. Different combinations of their effects (including associated common variants) have been evaluated in simulation studies resulting in different possible scenarios [9]. Rare and low-frequency variants were also estimated to contribute to a substantial part of heritability [25]; however, this hypothesis has not yet been validated in actual studies because their contribution remains limited so far [26-29]. A recent study on height, one of the largest to-date to explore the impact of rare and low-frequency variants on a complex trait, was able to identify rare variants with large effect sizes associated with the trait [30]. However, this study also revealed that they explained only a very small portion of heritability. They also observed a

positive association between minor allele frequency (MAF) and heritability meaning that common variants explain more of complex trait heritability than rare or low-frequency variants individually even if the latter harbor larger effect sizes [30]. Others have suggested that part of the missing heritability could also be explained by epistasis [19], parent of origin effect [31], epigenetics [32] or structural variants [21].

1.3 Assessing functional impact of variants

One way to better understand the genetic basis of complex traits is to link genetic variants to cellular traits (ex: gene expression, DNA methylation, histone modifications, etc.). In order to do that, lessons from past GWAS studies can help guide future genetic research: 1) over 90% of the GWAS hits reside in the noncoding part of the genome, 2) the associated loci highlight broad genomic regions as large as 100kb, 3) GWAS hits are mainly found in open and active chromatin identified by DNase I hypersensitive sites [17], and 4) a large portion of associated SNPs have an impact on gene expression levels [18]. Therefore, linking significantly or marginally associated SNPs to cellular and functional traits could help overcome the caveats of GWAS in three ways. First, these types of studies could help link non-coding variants to their gene of interest. Second, they could permit deciphering of their molecular effects. Finally, they could identify potentially interesting variants that do not reach genome-wide significance cut-offs. Another important aspect is that since these epigenetic features are cell type specific [17, 33, 34], they could help to better assess cell types that are implicated in the development of the diseases. Here I will present an overview on how *cis*-acting (i.e. acting locally) genetic/epigenetic interactions can be a useful tool to understand GWAS results and identify new genes and loci associated to complex traits

focusing on DNase I Hypersensitive sites (DHS), expression or methylation quantitative trait loci (eQTLs and mQTLs) and allele specific expression (ASE) or methylation (ASM).

1.3.1 DNase I Hypersensitive sites (DHS)

Regulatory regions of the genome can be identified through the use of the deoxyribonuclease I (DNase I) enzyme that preferentially targets open and active chromatin. DNase I hypersensitive sites (DHSs) have been used extensively to map regulatory DNA regions like enhancers, promoters, insulators, etc. [35]. DHSs were catalogued in a large number of cell types (around 350 cell types and tissues) by the ENCODE Project and the Roadmap Epigenomic program [36]. These efforts showed that GWAS hits are enriched in DHS and that they fall into the DHS of cell types or tissues relevant to the disease being studied [17]. DHS data also helped pinpoint the importance of certain cell types in specific diseases without considering previous knowledge of the disease pathophysiology [17, 37, 38]. Those results highlighted the importance of assessing *cis*-regulatory mechanisms in a diseased-linked, cell-type specific manner to better understand the functional aspect of GWAS hits.

1.3.2 Linking SNPs to gene expression: expression Quantitative Trait Loci (eQTL) and Allele-Specific Expression (ASE)

One way to identify *cis*-regulatory SNPs is to assess their impact on gene expression and help link them to the gene(s) of interest. It was previously used to understand the functional impact of GWAS hits in disease-relevant cell-types. The two general approaches that I am going to describe are quantitative trait loci (QTL) and allele-specific analyses (AS). The QTL approach measures the effect of a genetic variant on a functional aspect (like gene expression or DNA methylation) by correlating it to the different genotyping groups across individuals. The AS

approach measures the functional effect for each allele in a single individual at heterozygous sites. The great advantage of AS is that it requires smaller sample sizes than QTL. In fact, the *trans*-acting effect (distal effects) on *cis*-regulatory SNP can confound the results in QTL but not in AS studies since it is an intra-individual measure [39]. However, one drawback of AS is that it is usually based on next-generation sequencing data, which can result in greater costs compared with QTL studies that can be done using arrays.

The association of an allele with greater expression compared to another has been highly explored through eQTLs. Using expression arrays or RNA-sequencing, it allows linking a SNP to one or more genes located nearby. This method relies on the transcript abundance across samples. In the case of allele specific expression, it is based on the allelic imbalance measured in individuals' heterozygous sites lying within or close to transcripts and measures the relative expression between two allelic transcripts [40]. Since this is an intra-individual measure, bias coming from environmental factors and trans-genetic backgrounds are not confounding and the *cis* component can be directly measured. It was also previously shown that it is more sensitive compared to the QTL approach and that a 8-fold smaller sample size is needed to achieve similar power [41]. One drawback of ASE is that homozygous sites cannot be assessed but has interesting advantages like the control for *trans* effects. eQTLs and ASE mapping have been quite effective in retrieving functional and biological information from GWAS hits. First, GWAS hits are enriched for eQTLs [18]. They not only help pinpoint the gene of interest but they can also assess important cell types related to the disease development [42, 43].

1.3.3 Linking SNPs to DNA methylation: methylation Quantitative Trait Loci (mQTL) and Allele-Specific Methylation (ASM)

DNA methylation is the addition of a methyl group at the C5 position of the cytosine by the DNA methyltransferase enzyme and mostly happens in a CpG context. It is a heritable genetic mark and when occurring in canonical regulatory regions like promoters and enhancers, it can lead to disruption of the transcription process. DNA methylation in the gene body plays a role in preventing spurious transcription of the gene. Assessing the differential DNA methylation levels between cases and controls can be done using epigenome wide association studies (EWAS). It can lead to the identification of biomarkers for the disease that can be linked to either genetic factors (in *cis* or *trans*), environmental factors or even the disease itself.

The genetic influence on DNA methylation can be assessed using methylation QTLs (mQTLs) or allele-specific methylation (ASM). mQTL assesses the correlation between the methylation level at a CpG and genotype of a nearby SNP. ASM directly measures the methylation level of each allele in a heterozygous individual. Just like eQTLs and ASE, mQTLs and ASM harbor a large set of cell-type or tissue specific sites [34]. They can help better understand the functional impact of GWAS hits, but not necessarily link them to the gene of interest. In fact, a large number of mQTLs or ASM sites occurs in enhancer or insulator regions located distal to the gene. However, they can identify sequence elements important for the disease, which cannot be assessed by eQTLs or ASE. Also, only a small overlap was observed between mQTLs and eQTLs making them complementary rather than redundant [44]. Additionally, mQTLs and ASM can also help interpret EWAS data by aiding in differentiating the changes in DNA methylation attributed to genetic or environmental effects. Both ASM and mQTLs in specific cell-types were useful to identify new pathways and biological mechanisms linked to diseases [45-47].

1.3.4 Rare variants and cellular traits

The link between genetics and cellular traits was mostly explored for common variants and only a few studies assessed the impact of rare variants. They showed evidence of their potential impact on gene expression where an enrichment of rare variants was observed in the vicinity of genes at the extremities of the expression spectrum [48-50] and where the effect was heritable [51]. Only one study looked at their impact on DNA methylation and showed that collapsing rare and low-frequency variants together identified CpG methylation associated with a group of variants [52]. Even though they appear to be important in the regulation of gene expression, a lot of work is still needed to better delineate the functional impact of rare and low-frequency variants.

1.4 Exploring rare and low-frequency variants in complex traits

Following the GWAS era, genetic variants located in the rare (MAF <1%) and low-frequency (MAF 1-5%) spectrum were thought to explain part of the missing heritability of different complex traits. Simulation studies explored different scenarios of their implication independently or in combination with common variants of small effect sizes [9]. More and more studies have started to explore the impact of rare and low-frequency variants on complex traits. Even though they tend to explain a smaller part of complex trait heritability than expected from simulation studies, they appear to contribute to the architectures of these diseases [30]. In order to predict disease risks, it is important to identify these variants in the context of personalized/precision medicine. The promise of rare variants in understanding complex traits resides in their potentially easier interpretation. A large fraction of variants affecting protein function are rare, therefore associated coding variants that are identified are more likely to directly point to the

gene of interest. The rare variants are also usually not in linkage disequilibrium (LD) with multiple SNPs meaning that they can point to the exact region of interest. The identification of rare and low-frequency variants in complex traits can also confirm previously known loci or identify new biological pathways or gene of interest [30]. However, just like common variant studies, rare variants will need to be explored using large samples sets [53].

In this section of the thesis, I will describe different strategies to explore rare variants as well as what is known so far regarding their distribution across populations. Different approaches are used to assess low-frequency and rare variants: 1) custom genotyping arrays, 2) whole-exome (WES) or whole-genome sequencing (WGS) and 3) genotyping imputation.

1.4.1 Custom Genotyping arrays

The design of custom genotyping arrays to study rare variants usually focuses on specific diseases and previously identified target regions of interest. They typically target variants contained in haplotypes of interest identified through sequencing. One example is the Immunochip array designed to replicate and fine map loci from 12 autoimmune and inflammatory diseases [54]. It includes the top 3000 loci that were previously associated with each disease as well as all the known SNPs in the regions identified in the first version of the 1000 Genomes project (1KG) or resequencing initiatives. The purpose of the Immunochip was to identify true association and fine mapping of the loci for each disease as well as pleiotropic effects across the different diseases [54]. It is also more cost effective than traditional GWAS chip (around 80% less), allowing assessment of a larger number of samples to increase power [54]. However, it has a few limitations: 1) the Immunochip design was based on the first version of the 1KG pilot project, which has incomplete coverage [55], 2) is mainly restricted to European samples and 3) it relies on previous knowledge, thus newly identified loci are not as well covered

since it does not cover the whole genome [54]. Despite those limitations, the Immunochip allowed for the identification of new loci and helped better refine previously known ones associated with different autoimmune and inflammatory diseases like celiac disease [56], psoriasis [57], rheumatoid arthritis [58], multiple sclerosis [59] and inflammatory bowel disease (including crohn's disease and ulcerative colitis) [60]. Other examples are the Metabochip designed to study metabolic disease [61] and the ExomeChip, which includes mostly variants in the protein-coding regions of the genome [62].

1.4.2 Whole-exome and whole-genome sequencing

In the 2000s, the rapid development of new sequencing technologies led to decreased sequencing cost (https://www.genome.gov/sequencingcostsdata/), which in turn, resulted in the increasing use of WES and WGS to measure the association of rare and low-frequency variants. WES is a targeted approach that focuses on the coding region, which represents approximately 1.2% of the genome. It allowed getting higher coverage in a larger sample set at a cheaper cost than WGS. It also focused on a more easily interpretable part of the genome. Most of the WGS were performed at low depth impairing the accuracy of identified variants. However, the decreasing cost of sequencing will lead to WGS replacing WES, which will be a great asset based on what is known about the genetic architecture of complex traits. One of the first and most important works regarding this is the 1KG that began in 2008 and had the ambitious goal to sequence, using low-depth WGS and WES, over one thousand individuals from 14 diverse populations [55, 63]. This helped to catalog most of the common genetic variations and identify new rare ones. A few important lessons on rare and low-frequency variants have come out of the 1KG and other studies. They observed that the majority of variants are rare and population specific [63, 64]. A larger portion of low-frequency and rare variants are found in the coding regions compared to

common variants reflecting potential purifying selection effect [64]. Rare and low-frequency variants are also more functional and deleterious for protein coding genes compared to common variants [64]. Even though there is a high potential for identifying interesting rare and low-frequency coding variants in complex traits, non-coding alleles also appear to be interesting due to their enrichment in functional domains like transcription start sites (TSS) or DHS [65]. So far, the results obtained from WES and WGS reflected the population genetics model by observing an inverse relationship between the frequency of the allele and its effect size, echoing what was observed in Mendelian diseases residing at the very end of the spectrum [66] (see section "population genetic evidence" for more details).

A growing number of large-scale sequencing projects have explored the impact of rare and low-frequency variants on complex traits. Projects like UK10K [65], deCODE [67], SardiNIA [68] or GoNL [69] helped better understand variants implicated in both complex traits and population genetics. A more in-depth description of these studies on asthma and allergy related-traits is presented in the "Asthma and allergy related-traits" section.

1.4.3 Imputation

Imputation is a method used to statistically infer missing genotypes in a large population and is based on known genotypes from this population. Data obtained from WES and WGS can be used to impute the genotypes in a large sample set in order to increase power of the association test. This strategy is probably the most cost-effective of the three. It relies on available genome-wide genotyping data and reference panels that are available like the HapMap project [70, 71], which was the first available one. It was followed a few years later by the 1KG [63] and other efforts such as the UK10K Cohorts project and more recently the 100,000 Genome project in the UK, which assessed a large number of samples from British decent. Using the UK10K and 1KG panel

increased imputation accuracy at the low-frequency level (0.05%<MAF<5%) in the European populations [72]. However, they still remain limited regarding the imputation accuracy of rare variants (MAF<0.5%) [65, 73, 74]. Finally, the Haplotype Reference Consortium has put together all publicly available WGS data from 20 studies of European descent to create the largest reference panel [75]. Their goal was to create a large and diverse imputation panel that would allow for better results when imputing in samples with a genetically diverse background [75]. This panel should also allow for better imputation of low-frequency and rare variants in European samples.

Since low-frequency and, more strikingly, rare variants have arisen more recently, they are often restricted to specific populations and thus cluster geographically. To obtain a better imputation accuracy as well as to assess population specific variants, it is important to include samples from the population of interest. The importance of utilizing population specific panels for assessing rare variants has been shown in different sequencing studies [68, 69, 76] and this yields better accuracy than increasing the size of the reference panel (i.e. the number of haplotypes) [76].

1.4.4 Population genetic evidence

Human migration and rapid recent population growth have led to a large number of rare and low-frequency variants that are either population or individual specific [64]. The advent of next-generation sequencing revealed the excess of rare and low-frequency variants in the different human populations [64]. According to population genetic studies, rare coding variants are more deleterious and damaging compared to common variants due to purifying selection [29, 64]. In fact, most of the variants affecting protein-function identified to date are rare [64]. Since these variants have arisen more recently in the population, they had less time to be removed by the

evolutionary selection process. This may increase they probability to be related to disease development.

Since only a small number of rare variants are shared across populations/continents, they were examined more closely on their pattern in different populations. Some study stated that populations that underwent bottlenecks should be enriched in deleterious variants because of a reduction of the selection efficacy [77]. This could be even more pronounced in founder populations, which are genetically homogeneous populations usually due to demographic circumstances [78, 79]. However, other studies stated the opposite, where no enrichment of deleterious variants was observed, probably due to the short timeline where no accumulation was possible [80-82]. The latter question is still being debated but one advantage of the founder population would reside in the genetic drift resulting in the higher frequency of some variants private to the population.

1.5 Asthma and allergy related traits

Allergic diseases that comprise asthma, allergic rhinitis, atopic dermatitis and food allergies are a collection of diseases that are characterized by an immune-mediated inflammatory response to allergenic substances that are normally harmless. They have wide incidence variations from 2-4% in Asian countries and higher rates in developed countries, including Canada, with prevalence ranging from 15% to 20% for asthma [83-85], 10% to 40% for allergic rhinitis [83] and 1% to 20% for atopic dermatitis [86]. It is believed to affect from 300 to 500 million people around the world and could increase while more countries adopt a westernized lifestyle [83, 85]. Severe asthma has associated mortality: the number of deaths worldwide is estimated at around 250,000, which represents about 1 in every 250 deaths [83, 85]. It is also linked to an important

socio-economic cost linked to absenteeism, loss of productivity and emergency visit. All these elements point towards a need to better characterize disease subtypes, develop better treatments and personalized medicine.

1.5.1 Asthma and allergy pathophysiology

Asthma and allergic diseases are common complex traits that are often co-occurring in the same individual or families. They have an important environmental and genetic component making them complex and heterogeneous diseases. The "atopic march" is a process comprising sequential progression of allergic conditions usually leading to the development of asthma, allergic rhinitis or both [87, 88]. The presence of atopic dermatitis (eczema) combined with IgE modulated response to food or aeroallergen at a young age is usually the first clinical manifestation. Around 30% of children with atopic dermatitis will go on to develop asthma and more than 60% of them will develop more severe allergic disease later in life such as allergic rhinitis [89-91]. Also, a large majority of asthmatic patients also present allergic rhinitis (>80%) and 20% to 40% patients exhibiting allergic rhinitis also have asthma [92]. This indicates potentially shared biological mechanisms and pathways between these different clinical manifestations.

Asthma is a chronic inflammation of the airways characterized by airway obstruction, airway hyper responsiveness, lung remodeling, wheezing, and excessive mucus production. It is sometimes combined with allergic response in 80% of the affected children and 60% of the affected adults [93]. Monozygous twins show greater concordance than dizygotic twins. Its heritability, estimated from twin studies, ranges from 35 to 70% [94-96]. This can be explained by the fact that different exposures at different times during life can result in different risks of disease and age of onset. For example, exposures to bronchiolitis [97] or rhinovirus [98] in early

life, smoking or exposure to second-hand smoke [99] and occupational exposure (work environment [100]) can influence the development in the disease. Heritability is inversely correlated with age of onset of the disease, thus genetics plays a greater role in childhood-onset asthma [101]. The other asthma and allergy related traits and intermediate phenotypes also show significant heritability: 30% to 90% for allergic rhinitis [94, 102, 103], 70% to 85% for atopic dermatitis [102], 35% to 85% for serum IgE levels [104, 105] and 25% to 40% for eosinophil (Eos) counts [104, 105].

1.5.2 Genetics of asthma, allergy and other related diseases

As with any other complex trait, the approaches to study these diseases have evolved over time along with the arrival of new technologies. Starting from candidate gene and genome-wide linkage studies all the way to GWAS and sequencing studies, hundreds of genes have been associated with asthma, allergies, atopic dermatitis and allergic rhinitis. Despite the high clinical heterogeneity of the diseases and the importance of environmental exposure, GWAS identified many SNPs associated with the traits and were replicated across studies (summarized in Table 1). They either reinforced the importance of some genes that were already linked to the trait (ex: *IL33* in asthma or allergy, *FLG* in atopic dermatitis [106]) or identified new genes (the 17q12-21 locus *ORMDL3*, *GSDML* and *ZPBP2* in asthma). However, those variants explained very little of the heritability due to their small effect size.

Asthma definition remains difficult because it is seen as a plethora of similar diseases, thus making genetics studies very difficult. One way to overcome this challenge is to focus on intermediate quantitative phenotypes that can be measured more precisely like lung function, serum IgE levels, and blood Eos counts. These are part of asthma and allergy endophenotypes, which are clinical and biological markers that help define disease subtypes. Focusing on such

endophenotypes can help identify loci associated with asthma and allergic diseases or only the trait itself. However, one drawback is that these traits might also be linked to other disease like lung function for chronic obstructive pulmonary diseases (COPD).

To establish asthma diagnosis, the patient history is first assessed and usually confirmed by lung function test. The lung function is measured using spirometry which measures the forced vital capacity (FVC), defined as the amount of air that can be forcibly blown out of the lung when taking a deep breath, and the forced expiratory volume in one second (FEV_1) that measures the same thing but in the span of one second. The Tiffeneault index (FEV_1/FVC) is a common measure to assess the airway obstruction in lung disease such as asthma and COPD. So far, GWAS results of lung function poorly overlap with the asthma and allergy related traits and better results were obtained with COPD.

To assess the allergic aspect of the disease, skin prick test and measurement of serum IgE levels are usually used. IgE is one of the five Igs and is known to be an anaphylactic or allergic antibody. Strong correlation between IgE presence and asthma and allergy diagnosis and severity was previously observed [107, 108]. More than half of the GWAS hits of serum levels overlap with asthma, allergy, allergic rhinitis or atopic dermatitis (Table 1).

Eos are a cell type implicated in the initiation and propagation of inflammation and immune response in asthma, atopic dermatitis, allergic rhinitis and allergy [109]. Increased concentration in blood and tissues is a hallmark of certain forms of asthma (mostly allergic asthma and severe eosinophilic asthma) and is usually positively correlated with severity of the disease. They also potentially play a role in airway remodeling and are often the main inflammatory cell type present in the airways of asthmatic patients [110]. A few GWAS have assessed the impact of

genetics on Eos counts, but the results revealed a large overlap between the identified loci and previously identified asthma and allergy GWAS hits [111].

Finally, a GWAS tried to identify genetic factors underlying the atopic march [112] focusing on individuals presenting early-onset atopic dermatitis and childhood asthma. They observed a stronger contribution of atopic dermatitis genes compared to asthma suggesting their importance in the process.

Table 1: Summary of GWAS discovery for asthma and allergy-related traits

Trait	**Number of studies[1]**	**Number of genes / loci**	**Main replicated loci[2]**	**Reference**
Asthma	18	92	*ORMDL3, GSDMB, GSDMA, IL33, IL18R1, IL1RL1, HLA-DQA1*	[113-130]
Allergy	4	60	*HLA-DQA1, HLA-DQB1, HLA-C, C11orf30, GSDMB, IL1RL1, IL33, LPP, LRRC32, TLR1, TLR6, WDR36*	[114, 125, 131, 132]
Allergic rhinitis	3	19	*C11orf30, LRRC32*	[133-135]
Atopic dermatitis	9	106	*FLG, IL13, IL18RAP/IL18R1, KIF3A, RAD50, TNFRSF6B, OVOL1, C11orf30, ACTL9*	[136-144]
Intermediate phenotype	**Number of studies[1]**	**Number of genes / loci (overlap with main trait)**	**Main replicated loci[2]**	**Reference**
Serum IgE levels	4	16 (10)	*FCER1A*	[135, 145-147]
Eos	3	10 (3)	*GATA2*	[111, 148, 149]
Lung function[3]	8	33 (1)	*AGER, CDC123, HHIP, HTR4, GSTCD, THSD4*	[150-157]

Summary including only SNP p<5e-8. 1) studies have at least one SNP with p<5e-8. 2) Gene associated in at least two GWAS. 3) Lung function studies on COPD patients were removed.

1.5.3 Rare variants in asthma and allergies.

Torgerson et al. first explored the impact of rare variants in asthma where they performed a resequencing study of nine previously associated genes [158]. They observed an excess of rare variants in four genes contributing to the asthma phenotype, where the effect was predominantly

due to noncoding variants [158]. Another study used the ExomeChip to assess the impact of rare and low-frequency coding variants in three ethnic groups and identified one low-frequency variant associated with the trait in Latinos and three genes associated in gene-based tests [27]. Most of their findings were exclusive to a specific racial group, which was expected due to rare variations being private to certain ethnicities [27]. Other studies focusing on refined or intermediate phenotypes identified rare variants associated with the trait: bronchodilator response in asthmatics [159], asthma diagnosis following severe infection with respiratory syncytial virus [160], extreme lung function and airway obstruction [161].

There were also studies exploring rare and low-frequency variants in asthma and allergy intermediate phenotypes. A rare variant disrupting a canonical splice site of the *IL33* gene has been associated with reduced blood eosinophil counts and reduced risk of asthma in the Icelandic population and replicated in European populations [67]. Even though the Immunochip was used to identify SNPs associated with atopic dermatitis in multiple studies [137, 162], only one low-frequency variant in *PRR5L* was associated with the trait [162]. Finally, a resequencing study of interferon pathway genes identified a rare functional variant in *IFNGR1* gene associated with higher risk of eczema herpeticum in patients affected with atopic dermatitis [163].

So far, these studies showed the implication of rare and low-frequency variants in asthma and allergy related trait genetic architecture. Rare and low-frequency variants studies confirmed the importance of certain genes such as *ADRB2 [160], MTHFR* [27]*, GSDMB* [27]*, ZPBP2* [27], *FLG* [160], *NOS1* [160], *IL33* [67] and identified new ones like *GRASP* [27] and *PRRL5 [162]*, suggesting the importance of looking at the lower frequency spectrum of variants to better understand diseases.

1.5.4 Linking genetics to cellular trait to understand asthma and allergy related traits

Most of the studies linking GWAS hits to cellular function to understand asthma and allergy related traits have explored asthma and lung function phenotypes. They also only explored so far, the link between the SNP and gene expression through eQTLs in different cell types like lung tissue [164], airway epithelial cells [165], CD4+ lymphocytes [166], lymphoblastoid cells (LCL) [122], bronchial epithelial biopsy and bronchial alveolar lavage [167]. To date, no studies have explored eQTLs specifically in allergic rhinitis, atopic dermatitis, Eos counts or serum IgE levels.

Combining gene expression data to previously associated SNPs in asthma provided better elucidation of the biological function underlying these loci. When the well-known 17q21 locus was identified [122], the associated allele was first linked to *ORDML3* expression in LCLs from the affected children [122]. This observation was also confirmed in CD4+ lymphocytes, white blood cells and lung tissues, where other SNPs regulating the expression of *ORDML3* as well as *CRKRS, GSDMB* and *GSDMA*, were identified [164, 168-170]. However, the strongest eQTL of this region in lung tissues pointed to the *GSDMA* gene, which harbored an opposite effect compared to the other three genes [164]. Another example is the delineation of the known loci like *IL1RL1/IL18R1* [171, 172] and helped identify not only the gene of interest but also the tissue. In fact, eQTLs for these genes are present in lung, airway epithelial cells and distal lung parenchyma but not whole-blood, pointing towards increase risk of inflammation in the lung.

Other studies directly combined GWAS results with eQTL studies in different tissues and cell-types to identify new genes of interest [173-176]. For example, a gene-based approach combining 16 eQTLs studies and two asthma/allergy GWAS confirmed a set of previously

known genes and identified four new ones [174]. Two of the novel genes were shown to induce IL-33 release followed by eosinophil airway infiltration in mice [174].

1.6 Leveraging founder and isolated populations to study complex traits

Studies require larger sample sizes to reach sufficient statistical power and thus focus on larger heterogeneous populations. By combining different studies to leverage greater power, it introduces heterogeneity at the genetic level (difference in allele frequencies between populations) but also in environmental exposures, cultural habits (ex: life style, diet) and in disease diagnostics/classification, which could lead to reduced power. Population stratification is also an important issue, especially when studying rare and low-frequency variants, since they are usually private to a population thus making the effect stronger compared to common variants [63, 177]. Methods known to correct for it (for example principal component analysis (PCA) [178]) do not seem to work as well in the case of rare variants testing [179]. One strategy to reduce heterogeneity is the use of founder and isolated populations.

Founder and isolated population can provide a power boost to study rare and low-frequency variants [53]. They are usually derived from a subset of a population after a founding event: populating new territories, war, famine, environmental event, epidemic, etc. [180]. Usually after this founder event, a small number of individuals become isolated for a few generations due to geographic, cultural, religion, national identity or language reasons [181]. The founder populations arising from historic European migration show higher homogeneity and genetic drift comparatively to their population of origin. The nature and characteristic of the events also have an impact and may differentiate the level of homogeneity observed in a founder population: the number of founders and the number of bottleneck events, the duration of isolation, the population

growth (expansion) and the absence of immigration from neighboring populations (gene flow) [180].

The homogeneity of founder populations is a great advantage in the study of rare and low-frequency variants in complex traits. Some variants implicated in the disease risk could reach higher frequencies compared to outbred populations due to bottlenecks, genetic drift, adaptation and selection [68], facilitating their identification. One example of this resides in the increased incidence of certain recessive and rare disorders in founder populations [182]. Other studies also identified private variants associated with complex traits like type 2 diabetes [183], height [184] or lipid traits [185]. These variants were identified because they reached higher frequencies in the founder populations and could be important for personalized medicine in these populations.

In this thesis, we are using the Saguenay–Lac-Saint-Jean (SLSJ) asthma familial cohort [186]. This population is located in north-eastern Quebec and is known for its unique demographic history and founder effect, characterized by several population bottlenecks followed by rapid expansion. At the beginning of the 17th century, around 8,500 settlers came to the Vallée du St-Laurent from France. These people represent a large part of the 6 million French Canadian ancestries that now live in the province of Quebec. A subset of them migrated to the Charlevoix region where there was a rapid expansion after their settlement due to high birth rate. Because of overpopulation, another subset migrated to the SLSJ in the mid 1800's. The latter now represent 75% of the SLSJ founders thus had a reduced contribution of new immigrant after the first settlers. They became genetically isolated from France and at a lesser extent from the other regions in the province. Regional clustering of multiple hereditary diseases was observed [182].

The SLSJ asthma familial cohort has contributed to the understanding of asthma and allergic diseases in many ways. Candidate genes studies performed on this population identified genes

associated with the traits in the *IL1R2* gene pathway [187], the well-known 17q21 locus [187], the Vitamin D pathway [112, 188] and other genes like *CX3CR1 [189]*, *PLAU [190]* and *ATG5 [191]*. Subsequent studies focused on assessing the epigenetic mechanisms that could explain the link between previously associated genes to the diseases. Higher DNA methylation levels were observed in the *IL1R2* promoter region in asthmatics and allergic individuals in whole blood [192]. Other studies helped decipher the 17q21 locus [193-196]. Two GWAS were published using this cohort, one exploring different traits in the cohort alone [186] and the other being part of the larger GABRIEL consortium for a meta-analysis [121]. The cohort is also included in the Translational Asthma Genetic Consortium (TAGC) consortium GWAS that will be released in 2017. Finally, the SLSJ asthma familial cohort was also used in an EWAS to assess the link between methylation patterns and serum IgE levels in peripheral blood as well as Eos [197].

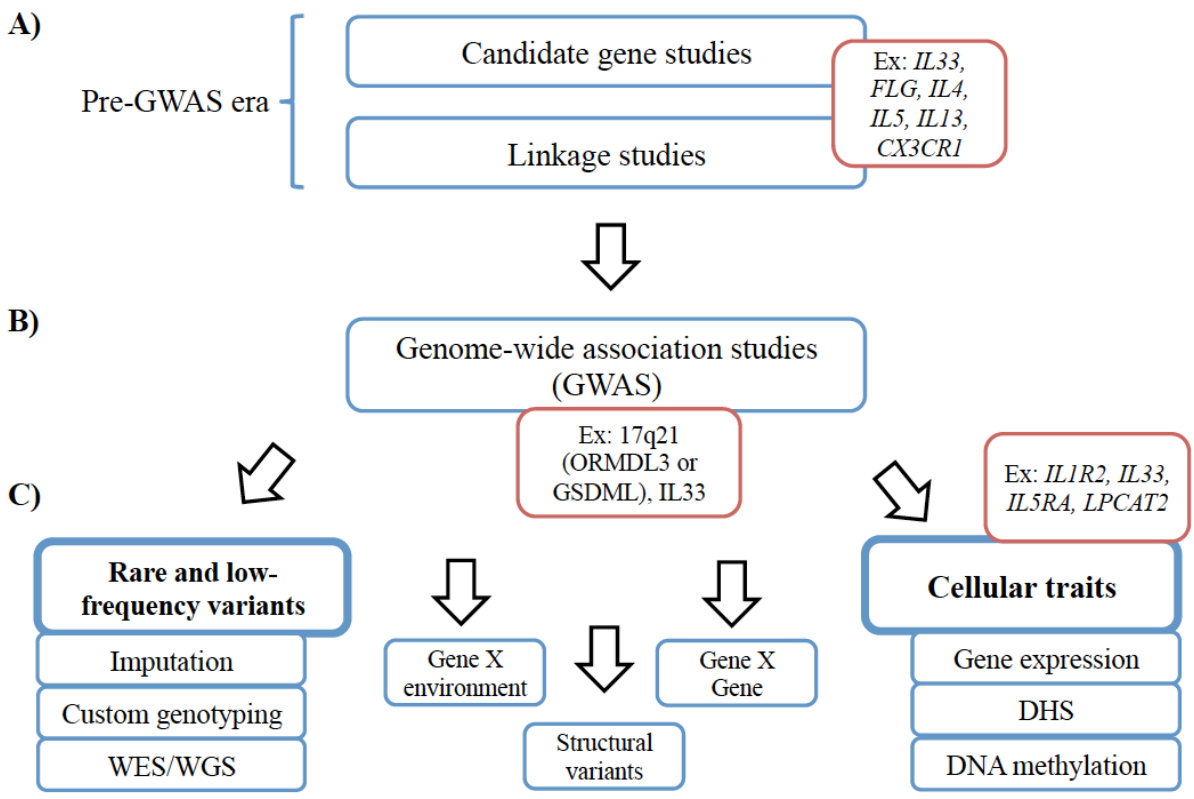


Figure 1: Representation of pre and post-GWAS era genetic approaches to study complex traits. A) Candidate genes and linkage studies were one of the first ways to study the genetic aspect of complex traits. A few genes were identified in the SLSJ asthma cohort using one of the techniques and examples are listed in the red boxes. B) These studies were followed by the advent of Genome-wide association studies (GWAS). The SLSJ asthma familial collection took part of large consortium that identified loci that were highly replicated. Examples of these loci are listed in the red box. C) New strategies were developed to complement GWAS findings. Two of them (in bold) are explored in this thesis: assessment of rare and low-frequency variants and linking GWAS hits to cellular traits. Different ways to explore these two strategies are also listed as well as examples of genes that were identified in the SLSJ asthma familial cohort (red boxes).

1.7 Rationale, objectives and hypothesis

The objective of my doctoral thesis was to use different strategies to understand the genetic basis of complex traits and more precisely, asthma and allergy related-traits. Lessons from GWAS guided the research community to develop new strategies to fill in the blanks left unanswered, now that we know their important caveats.

We first designed a custom capture panel (Immune-genetics sequencing) that targets both coding and non-coding regulatory regions of immune cells. We wanted to develop a cost-effective way to study rare and low-frequency variants in autoimmune and inflammatory complex traits. The goals of this study were to 1) define interesting and functional regions to target that are properly suited to the diseases we would like to study, 2) assess the functional impact of the rare and novel variants identified using our custom capture panel followed by next-generation sequencing in healthy subjects, and 3) determine the impact of the newly identified variants on gene expression.

We used our Immune-genetics sequencing in our second paper to explore the impact of rare and low-frequency variants on asthma and allergy related traits in a founder population. We sequenced 149 trios from the SLSJ asthma cohort using our custom capture panel. We first assessed the rare and low-frequency variants distribution in this founder population compared to four European populations, including Finland, which also has an important founder effect. We then assessed the impact of rare and low-frequency variants on lung function, serum IgE levels and eosinophil counts.

Finally, the third paper aimed to identify new genes associated with allergic rhinitis with or without asthma in the SLSJ asthma familial cohort. We performed a GWAS and an EWAS and

combined marginally associated SNPs and CpGs using mQTLs to identify new genes associated with the trait.

Chapter 2

Preface: Bridging Text between Chapters 1 and 2

One of the strategies to better understand the genetic basis of complex traits that is becoming more and more popular over the years is the exploration rare and low-frequency variants. Whole-exome sequencing (WES) has been the most popular due to its lower cost compared to whole-genome sequencing (WGS). However, WES focuses only on the coding portion of the genome and does not explore any non-coding regions, where genome-wide association study identified the majority of the variants. We wanted to develop our own custom capture panel that targets both coding and non-coding region to assess the impact of rare and low-frequency variants in autoimmune and inflammatory diseases at lower cost compared to WGS. In this chapter, we described how we designed our capture panel and how the variants identified with it are potentially highly functional.

The word “Immunoseq” was replaced by “Immune-genetics sequencing” in the following manuscript due to trademark concerns.

Chapter 2: Immune-genetics sequencing: the identification of functionally relevant variants through targeted capture and sequencing of active regulatory regions in human immune cells

Authors:
Andréanne Morin[1,2], Tony Kwan[1,2], Bing Ge[2], Louis Letourneau[2], Maria Ban[3], Karolina Tandre[4], Maxime Caron[2], Johanna K Sandling[4,5], Jonas Carlsson[5], Guillaume Bourque[1,2], Catherine Laprise[6], Alexandre Montpetit[2], Ann-Christine Syvanen[5], Lars Ronnblom[4], Stephen J Sawcer[3], Mark G Lathrop[1,2], Tomi Pastinen[1,2,*]

Affiliation
1 Department of Human Genetics, McGill University, Montréal, Quebec, Canada
2 McGill University and Genome Québec Innovation Centre, Montréal, Quebec, Canada
3 University of Cambridge, Department of Clinical Neurosciences, Cambridge, United Kingdom
4 Department of Medical Sciences, Section of Rheumatology, Uppsala University, Uppsala, Sweden
5 Department of Medical Sciences, Molecular Medicine and Science for Life Laboratory, Uppsala University, Uppsala, Sweden
6 Département des sciences fondamentales, Université du Québec à Chicoutimi Saguenay, Québec, Canada

*Corresponding author: Tomi Pastinen:
Andréanne Morin: andreanne.morin@mail.mcgill.ca
Tony Kwan: tony.kwan@mcgill.ca
Bing Ge: bing.ge@mail.mcgill.ca
Louis Letourneau: lletourn49@gmail.com
Maria Ban: mb531@medschl.cam.ac.uk
Karolina Tandre: karolina.tandre@medsci.uu.se
Maxime Caron: mcmax135@gmail.com
Johanna K Sandling: johanna.sandling@medsci.uu.se
Jonas Carlsson: jonas.carlsson@medsci.uu.se
Guillaume Bourque: guil.bourque@mcgill.ca
Catherine Laprise: Catherine_Laprise@uqac.ca
Alexandre Montpetit: alexandre.montpetit@mail.mcgill.ca
Ann-Christine Syvänen: ann-christine.syvanen@medsci.uu.se
Lars Rönnblom: lars.ronnblom@medsci.uu.se
Stephen J Sawcer: sjs1016@cam.ac.uk
Mark G Lathrop: mark.lathrop@mcgill.ca
Tomi Pastinen: tomi.pastinen@mcgill.ca

Published in:
BMC Med Genomics. 2016 Sep 13;9(1):59. doi: 10.1186/s12920-016-0220-7.

2.1 Abstract

Background: The observation that the genetic variants identified in genome-wide association studies (GWAS) frequently lie in non-coding regions of the genome that contain *cis*-regulatory elements suggests that altered gene expression underlies the development of many complex traits. In order to efficiently make a comprehensive assessment of the impact of non-coding genetic variation in immune related diseases we emulated the whole-exome sequencing paradigm and developed a custom capture panel for the known DNase I hypersensitive site (DHS) in immune cells – "Immune-genetics sequencing".

Results: We performed Immune-genetics sequencing in 30 healthy individuals where we had existing transcriptome data from T cells. We identified a large number of novel non-coding variants in these samples. Relying on allele specific expression measurements, we also showed that our selected capture regions are enriched for functional variants that have an impact on differential allelic gene expression. The results from a replication set with 180 samples confirmed our observations.

Conclusions: We show that Immune-genetics sequencing is a powerful approach to detect novel rare variants in regulatory regions. We also demonstrate that these novel variants have a potential functional role in immune cells.

Keywords

Rare variants, immune disease, gene expression, next-generation sequencing, capture

2.2 Background

Genome-wide association studies (GWAS) have identified thousands of associated single nucleotide polymorphisms (SNPs) in hundreds of complex diseases [8] and have thereby provided unprecedented insights into the genetic architecture underlying these conditions [198]. However, because GWAS are inherently dependent upon there being meaningful linkage disequilibrium (LD) between relevant variation and the few hundred thousand common variants that are actually genotyped this method has limited ability to accurately assess the role of rare variants[199] and effectively only screens common variation [200]. This limitation has been suggested to contribute to the notable gap between observed heritability and that explained by the currently identified common variants - the so-called missing heritability [14]. Direct assessment of all variation through the next-generation sequencing of the whole genome would provide a comprehensive assessment that would necessarily avoid any dependency on LD but unfortunately remains prohibitively expensive. On the other hand, the targeted capture of genomic regions with high prior probability of containing relevant variation allows next-generation sequencing efforts to be focused and therefore substantially more affordable. This logic underlies whole exome sequencing which allows comprehensive assessment of coding variation and has enabled the identification of rare coding variants exerting large effects in a number of complex diseases [201-203]. It is notable that the majority of the associated variants identified through immune disease GWAS are located in non-coding regions of the genome that are enriched for regulatory elements that are active in immune cell types [17, 204, 205], suggesting that a resequencing effort focused in these regulatory regions would provide a highly efficient means to identify both common and rare variation of relevance in such diseases.

Using deoxyribonuclease I (DNase I) based sequencing (DNase-seq) international collaborative efforts such as the ENCODE [206] and NIH Roadmap Epigenomics [207] projects have established comprehensive maps of DNase I hypersensitive sites (DHSs) in multiple cell types. Sites which are markedly enriched for *cis*-regulatory elements active in those cell types such as enhancers and promoters [35, 36], show very high concordance with chromatin immunoprecipitation sequencing of histone marks for active enhancers or promoters [208, 209] and are enriched for SNPs (eSNPs) that influence the expression of local genes that show variable expression (expression quantitative trait loci, eQTLs) [208, 210, 211]. It has been noted that the enrichment of eSNPs is most pronounced in those functional elements that are located closest to their respective eQTL [209] and that there might be an inverse relationship between the effect size of *cis*-eQTLs and the minor allele frequency (MAF) of the relevant eSNP; suggesting that rare variants might have a higher impact on gene expression than common variants [48, 49, 212, 213].

Based on the overwhelming evidence from GWAS that common variants associated with immune disease likely influence disease risk by perturbing the regulation of gene expression together with emerging evidence indicating the existence of rare "high-impact" non-coding variation, we designed a custom capture panel, relying on contemporary regulatory element maps, to enable the targeted re-sequencing of immune regulatory regions - "Immune-genetics sequencing". Immune-genetics sequencing is designed to allow efficient re-sequencing of regulatory regions of relevance in immune cells (coding and non-coding) and thus enable a comprehensive assessment of all potentially relevant variation in these regions, both common and rare. The panel includes SNPs previously associated with immune traits as well as established immune cell eSNPs. Using Immune-genetics sequencing in parallel with

transcriptome sequencing (RNA-seq), we show that, after accounting for effects attributable to associated common variants, there are significant effects attributable to rare variants, and that these explain up to 14% of residual variation. Our results confirm that targeted capture and re-sequencing of regulatory regions active in relevant cell types provides an efficient means to identify rare variants of relevance in immune disease.

2.3 Methods

Design of the Immune-genetics sequencing custom capture panel:

We selected regulatory regions of immune cells using genome-wide DHS data from the ENCODE [206] and NIH Roadmap Epigenomics [214] projects. Data from twelve different immune cell types were utilized: CD3+, CD3+ cord blood, CD4+, CD8+, CD14+, CD19+, CD20+, CD34+, CD56+, Th1, Th2, Th17 (S1 Table). The entire genome was divided in 100bp bins and the DHS signals were normalized by calculating the number of reads per bin divided by the total number of reads. In each sample, the signals were ranked and the top 300,000 bins (representing the top 1% of the genome) identified, within each cell type bins were retained if they were identified in at least 50% of the available samples. For those cell types where only two samples were available, the selected bins were required to be present in both samples; in those cell types where only one sample was available (Th2, Th17 and CD20) all 300,000 bins were retained. The 100bp bins were then grouped into blocks of 50,000 bins each (i.e. 0 to 50,000 top bins, 50,000 to 100,000 top bins etc.) and when the overlap between sample blocks (from the same cell type) dropped below 50%, the blocks were eliminated. S1 Table shows the number of bins used and the number of samples available for each cell type. All selected regions were combined and bins were removed when at least 50% of a bin overlapped with an exome capture

region (SeqCapEZ Exome V3 Capture, Roche, 64.1Mb). Non-coding regions targeted by our design cover a total of 67.3Mb. The Immune-genetics sequencing custom capture was complemented by exome (SeqCapEZ Exome V3 Capture, Roche, 64.1Mb) and Human Leukocyte Antigen (HLA) regions (SeqCap EZ design, Human MHC design from Roche, 4.97Mb) totaling 138Mb for the panel.

Enrichment of GWAS hits in DHSs selected for the Immune-genetics sequencing custom capture panel design

GWAS hits were obtained from the National Human Genome Research Institute (NHGRI) (https://www.genome.gov/26525384, January 29th 2015). We selected SNPs from different disease categories: Immune and chronic inflammatory diseases (724 SNPs), associated to more than one immune or chronic inflammatory diseases (49 SNPs), Neuropsychiatric disease (65 SNPs) and Cancer (393 SNPs), including SNPs in LD using HaploReg V2 (r^2>0.9) [215]. Functional variants were selected from Monocyte and B-cell *cis*-eQTLs identified in the paper by Fairfax and colleagues [210]. Associated eQTLs with empirical p<0.001 after 1,000 permutations, for each top hit per transcript were retained for each cell type. ImmunoChip hits (224 SNPs) from five immune and chronic inflammatory disease studies [56-59, 216] were used. The analysis of overlap between Immune-genetics sequencing regions and SNPs was determined using bedtools (v2.17.0).

We compared the enrichment of GWAS or ImmunoChip hits and functional variants in DHS regions included in the Immune-genetics sequencing to other regions: 1) DHS from other cell types (S2 Table) selected in the same way as for the immune cells in the Immune-genetics sequencing design, 2) Same as in 1) but keeping only regions that do not overlap immune cell

DHS regions selected for Immune-genetics sequencing (compared to Immune-genetics sequencing DHS regions not overlapping with the other cell types' DHS regions), 3) an equal number of bins as in the Immune-genetics sequencing DHSs selected randomly from the whole genome in 1,000 iterations and 4) an equal number of bins as in the Immune-genetics sequencing DHSs selected randomly from the non-coding genome in 1,000 iterations. For the randomly selected regions, the whole genome was split into 100bp bins and 67,300 of them were selected, 1,000 times. Fisher's exact test was performed to evaluate the significance of the enrichment.

Design of the second version of Immune-genetics sequencing

Using coverage statistics from the first version of the Immune-genetics sequencing panel, we flagged poorly covered regions (<0.1X across all samples) or unusually high coverage regions (>120x across all samples), as well as ENCODE Blacklist regions for removal, and used the remaining regions to begin designing a 2nd version of our Immune-genetics sequencing panel. Additional regions totalling 7.243 Mb based on Digital Genomic Footprinting (DGF) data from ENCODE for CD4+, CD8+, CD19+ and CD56+ were added for this new panel.

Capture and sequencing

Thirty samples from the Swedish Uppsala Bioresource cohort were used as the discovery sample set in this study. The regional ethical review board in Uppsala, Sweden approved the study and all participants gave their informed consent. The Cambridge Multiple Sclerosis (MS) sample set was used as a replication set in this study. Eighty-six affected and 94 healthy controls were included for a total of 180 samples. DNA was prepared from Peripheral Blood Mononuclear Cells using standard methods. DNA quantification was performed using PicoGreen.

Whole-genome library preparation was performed using 500-1000ng of genomic DNA. Covaris focused-ultrasonicator E210 was used for shearing DNA into 150-1500bp fragments. LabChip EZ reader was used for fragment size evaluation and size selection was performed when needed. Libraries were prepared using the KAPA High Throughput (HTP) Library Preparation Kit (KAPA Biosystems). The end repair to produce blunt-ended double stranded DNA, adenylation of the 3'-ends, adapter ligation and amplification were performed following the recommendations from the kit manufacturer and cleaned using AMPure XP beads. The libraries were analyzed on LabChip and quantified using PicoGreen. Samples were then pooled (2X, 5X or 6X) using a total of 1 μg of library, followed by Roche NimbleGen SeqCap EZ Library instructions for the hybridization of the baits and the capture steps. The final amplification was done using KAPA HTP. Concentration, size distribution, and quality of the amplified capture were assessed using LabChip. Captured products were sequenced on the Illumina HiSeq2500 or HiSeq2000 with 100bp paired-end reads. The discovery sample set was captured with the first version of the panel, and the replication set was captured with the second version of the panel. For the second panel, the library preparation and capture steps were automated and performed using the Biomek FX (Beckman Coulter).

Read mapping and variant calling

Reads were aligned to Genome Reference Consortium Human genome build 37 (GRCh37) using bwa 0.7.6a. and variants were called using HaplotypeCaller v3.2 (GATK).

Variants quality control/SNVs validation

Quality cut-off was set at read depth ≥10, genotyping quality (gq) ≥70, and mapping quality (MQ) ≥50. These cut-offs were selected based on the comparison of the sequencing and genotyping data (Human Omni2.5 BeadChip in the 30-sample cohort or Human Omni5 BeadChip in the 180-sample set), available for all samples, where both had concordance of over 95% (S1 Fig). Indels were not included in our analysis.

To test the variant capture efficiency of Immune-genetics sequencing, we applied our panel to a Yoruban sample (NA18502) that has been sequenced at high depth by Complete Genomics [217]. We compared the accuracy of the heterozygous variants identified by Complete Genomics that overlapped with the panel regions with the variants identified using our custom capture panel (S2 Fig). DNA sequencing data from the NA18502 sample was downloaded from the public genome data repository (ftp2.completegenomics.com, assembly software version 1.10).

Annotation of variants

The GERP++ score was used as a metric for conservation to identify selectively constrained variants (http://mendel.stanford.edu/SidowLab/downloads/gerp/) [218]. We also used the CADD tool to score the deleteriousness of the identified variants (http://cadd.gs.washington.edu/) [219]. Coding variants were annotated using snpEff [220]. Common variants are defined as having MAF>=1% and rare variants are defined as having MAF<1% based on the allele frequencies from the 1000 Genomes Project [63]. Novel variants were defined as variants not observed in the 1000 Genomes Project or dbSNP141.

Shared vs cell-type specific DHSs

The DHS sets selected for each cell type were intersected to determine which bins are observed in all selected cell types or in a subset of the cells. Enrichment was measured by comparing the number of rare and/or novel variants to the number of common variants falling in each category of DHSs and the total observed in DHSs.

Identifications of variants that disrupt or create motifs

Each identified variant was tested for the impact of the reference and the alternate allele on transcription factor motifs ± 15 nucleotides from the variant position. Matrices for TRANSFAC (version 2009.4) were used with the Finding Individual Motif Occurrence (FIMO) scanning software, version 4.10.1, using a $p<1.42e-7$ threshold (Bonferroni correction: 0.05 / 351,088 SNPs =1.42e-7). Only motifs directly overlapping a variant were kept. A motif was considered as created if it had a significant matrix affinity score only with the alternate allele, whereas it was considered disrupted if it had a significant matrix affinity score only with the reference allele.

RNA-sequencing and allele-specific expression mapping

Purified T cells were isolated from the discovery set samples (eight CD3+ and 20 CD4+). RNA was isolated with miRNeasy Mini Kit (Qiagen) and 500ng of RNA was used to prepare libraries using Illumina TruSeq Stranded Total RNA Sample preparation kit following the manufacturer's instructions. Quality control was performed using Agilent Bioanalyzer and samples were sequenced on Illumina HiSeq2000 with 100bp paired-end reads. Raw reads were trimmed (quality: phred33≥30 and length n≥32), adapters were removed (using Trimmomatic V.0.32

[221]) and reads were aligned to the hg19 human reference (Tophat v.2.0.10 [222] and bowtie v.2.1.0 [223]) for 81.9% of the reads aligned. For the replication set, purified T-cell (CD4+ and CD8+) subpopulations were isolated from 180 subjects (86 multiple sclerosis patients and 94 healthy controls) for 73% of the reads aligned. For details see Lemire et al [224].
Allele counts were measured using the SNPs from Illumina Human Omni2.5 BeadChip (30 samples cohort) or Human Omni5 BeadChip (180 samples cohort) and imputation (1,000 Genomes Project, using the IMPUTE2 software). Haplotype phasing was performed using the SHAPEIT V2 software and allele specific expression was calculated using reads from whole genes as previously described [211]. We used the Allele-specific expression (ASE) association data calculated with the replication cohort for the first cohort because of the lack of power due to the small samples number. Since CD3+ cells were not assessed in the replication cohort, we use the combination of CD4+ and CD8+ data to get association p-values for this cell type.
Transcripts with association p-value <1e-5 were kept, and isoforms were removed based on normalized read counts for each gene (keeping the best covered isoform). A total of 3,859 transcripts for CD3+ cells and 3,428 transcripts for CD4+ cells in the 30 samples discovery set, and 5,536 transcripts for CD4+ and 5,594 transcripts for CD8+ cells in the replication set were included in the analysis.

Enrichment of rare variants in vicinity of allelically imbalanced (AI) genes

The fold difference between the expressed alleles was calculated as counts for the most abundant allele divided by counts for the less abundant allele. Thus, a fold difference of one corresponds to alleles that are expressed equally. Genes with fold difference between 2 and 9 were considered as

having allelic imbalance (AI). Genes with > 9-fold were considered to be enriched for imprinted loci or artefacts and were thus removed from the analyses.

We performed enrichment analysis for variants in DHS +/-20kb from each gene. We calculated the enrichment of rare variants in highly AI genes (ASE effect size between 2 and 9, 1 meaning both alleles are expressed equally) by dividing the proportion of AI genes with rare variants in correlated DHSs by the proportion of all tested genes with rare variants in correlated DHSs.

DNase –sensitive regions correlated to transcript promoters

NIH ENCODE Roadmap DHS datasets (n=317) were retrieved and binned into 100bp segments as described above. Using transcripts from GENCODE v15, we extracted all promoter regions (defined as transcription start site (TSS) +/-500bp). Across all of the DHS datasets, we correlated the normalized bin scores for these promoter region bins with all DHS bins +/- 1Mb.

Hi-C region linked with promoter regions

Hi-C data from GM12787 lymphoblastoid cell line were obtained from Rao *et al*. [225] (Gene Expression Omnibus accession number: GSE63525). We extracted all regions that overlapped promoter regions (1500bp from TSS) of gene where expression data was available, as well as the linked regions.

2.4 Results

Design of the Immune-genetics sequencing custom capture panel

In order to select the most relevant non-coding regions to target, we used DNase I mapping data available from the ENCODE and Roadmap epigenomics projects from 12 different cell types (CD3+, CD3+ cord blood, CD4+, CD8+, CD14+, CD19+, CD20+, CD34+, CD56+, Th1, Th2 and Th17, S1 Table) [206, 207]. The whole genome was divided into 100 base pair bins, which were ranked according to the DHS signal for all samples available for every immune cell type (Methods). The top 300,000 signal intensity bins for every cell sample from the ENCODE and Roadmap epigenomics project were used for the design of the Immune-genetics sequencing capture panel. The bins that were kept were required to be consistent in most (>50%) biological replicates used for each cell type. S1 Table shows the number of DHS signal intensity bins used and the number of samples available for every cell type. We combined these putative regulatory regions (67.3Mb) with the coding regions from exome capture and the HLA region. However, given the unique and complex role of HLA in immune disease risk along with the extreme sequence diversity of human Major Histocompatibility Complex, we exclude its analysis in the following discussion. Altogether, Immune-genetics sequencing covers a total 138Mb of the genome.

The Immune-genetics sequencing regions are enriched in pertinent GWAS hits and eQTLs

We estimated the sensitivity of this panel by determining the extent to which it captured known autoimmune and chronic inflammatory diseases associated SNPs listed in the National Human Genome Research Institute (NHGRI) GWAS catalogue ($p<5x10^{-8}$) [8]; or SNPs in high linkage disequilibrium ($r^2>0.9$) with these [215] (S1-S2 Table). We repeated this process using cancer

and neuropsychiatric diseases associated SNPs listed in the GWAS catalogue (assuming that immune cells play a less significant role in these conditions, although it some case, it can play one) and using *cis*-eQTL data for monocytes (CD14+) and B-cells (CD19+) from Fairfax *et al.* [210].

This panel includes SNPs in high LD (r^2>0.9) with 62% (448 SNPs) of the autoimmune disease associated variants listed in the GWAS catalogue (Fig 1A), 63% (140 SNPs) of the associated variants identified in key ImmunoChip studies [56-59, 216] (S3A Fig) and more than 68% (378 SNPs) of the eSNPs identified by Fairfax et al (Fig 1B) [210]. These observations indicate the potential of our design to identify variants associated to autoimmune disease as well as other variants with potential functional impact on immune cell function. In contrast, alternate panels based on DHSs from randomly selected tissues, or random genomic regions show significantly poorer performance (Fig 1C-D, S3B Fig).

Functional potential of rare and novel variants identified using Immune-genetics sequencing

Performing Immune-genetics sequencing on DNA from 30 healthy blood donors (Table 1) at a mean sequencing coverage of 52x, we found that on average 88% of the reads were located on or near target, >98% of the target regions were covered (only 1.90% of the bases were missing) and 95% of the target regions were covered by at least two reads.

Taking advantage of the high sequencing depth, we were able to identify rare and novel variants at high confidence. We defined rare variants as those having a MAF <1% in 1000 Genomes Project data (Phase3) and novel variants that have not previously been identified by either 1000 Genomes Project or dbSNP141. A total of 351,088 variants were identified, of which 275,042 were common, 50,004 were rare and 26,042 were novel (Table 2, S3 and S4 Table).

Comparing non-coding with coding variants we found a significantly higher proportion were novel (p-value=2.87e-175) and selectively constrained variants based on Genomic Evolutionary Rate Profiling (GERP++ ≥1 p-value=3.57e-60 and GERP++≥2 p-value=3.06e-47) (Fig 2A). Using GERP++ [218] and Combined Annotation Dependant Depletion (CADD) scores [219], we also observed that the proportion of selectively constrained variants was greater amongst the novel and rare variants than amongst the common variants (Fig 2B).

We next partitioned the variants called according to whether the DHS used in the design was shared among cell types or unique to one cell type. It has been previously shown that cell-type specific DHSs mostly overlap gene bodies and intergenic regions, whereas DHSs that are shared between cell types overlap with more active regions and promoters [226]. We observed a higher proportion of novel and rare variants compared to common variants in DHSs that are shared between cell types, compared to the ones that are unique for a single cell type (Fig 2C). A clear increase in enrichment is observed when variants present at cell type unique DHSs and variants that are in DHSs shared between two to twelve cell types are compared, with linear regression p-values of 1.35e-05, 2.41e-06 and 5.81e-05 for rare, novel and combined rare and novel variants, respectively. These findings indicate that rare and novel variants are enriched in more active genomic regions compared to common variants.

To further investigate the potential functional impact of the rare and novel variants in DHSs, we explored the proportion of variants that disrupt or create transcription-factor motifs, compared to common variants and GWAS hits (Methods). In comparison to common variants a significantly higher proportion of novel variants create (p-value=1.56e-38) or disrupt (p-value=2.43e-11) transcription factor motifs (Fig 2D). Rare variants show a slightly lower, but still significant

enrichment for created motifs than do common variants (Fig 2D, p-values for disrupted motifs = 0.16 and created motifs = 2.77e-07).

The functional impact on gene expression by variants identified using Immune-genetics sequencing

Given that rare non-coding variants in regulatory genomic regions can exert large *cis*-eQTLs effects and demonstrate extreme allele specific expression (ASE) bias [48, 49] we assessed the extent to which the rare and novel variants identified using Immune-genetics sequencing influenced gene expression in a second independent set of samples; T cells (both CD4+ and CD8+) from 180 individuals used in a parallel effort to map common SNPs resulting in ASE (Ban, Ge et al. manuscript in preparation), almost 400 RNA-seq datasets in total.

We also generated deep RNA-seq data from fractionated T cells (CD3+ or CD4+) obtained from the 30 individuals used initially. These data generated equivalent results, which are shown in the Supplementary materials (S4 to S10 Fig).

For each gene we counted and characterised (coding/non-coding and novel/rare/common) the variants lying in the immediate vicinity (gene +/-20kb) and determined the allelic imbalance (AI) in expression observed in each transcript. After adjusting for the average number of SNPs used to calculate AI for each transcript, we observed a higher proportion of transcripts with non-coding variants in their vicinity for transcripts where the higher AI level is independent of a common, rare or novel regulatory variant (S11-S12 Fig). A distinct increase in the proportion of variants was observed by comparing equally expressed transcripts with a <1.5-fold difference in their allelic expression with transcripts displaying AI with an ≥1.5, ≥2, ≥2.5, ≥3 and ≥3.5-fold

difference in allelic expression (Fig 3A). The increase in AI is more pronounced for transcripts flanked by rare or novel variants than by common variants. In order to control for the influence of common variants we repeated this analysis focusing on just those genes which are known to undergo ASE (Ban, Ge, et al. manuscript in preparation) and for which we had already mapped the common SNP contribution to *cis*-regulation by ASE-mapping [40]. This approach allowed us to include just those individuals that are homozygous for the relevant common eSNP and thereby exclude the influence of these common variants (S13 Fig). The same trend was observed for such transcripts when analysis was based exclusively on data from individuals homozygous for the local established common variant eSNP (S14 Fig).

This observation was then confirmed when rare and novel variants are considered together and this situation is even more pronounced when focusing on individuals homozygous for the relevant eSNP (Fig 3B). Rare and novel variants located in DHSs that are correlated to the transcript promoters are highly enriched transcripts with substantial AI (>=2 fold) compared to common variants, especially in transcripts with homozygous common eSNP (Fig 3B). The stronger the correlation between a promoter and a DHS is, the more it is enriched in rare and novel variants (p-value=0.0196), and is even stronger when looking at transcripts with homozygous common eSNP (p-value= 0.0024, Fig 3B). We also observed that the transcripts displaying higher AI show more enrichment for rare and novel variants in its vicinity, compared to common variants (Fig 3C). This was also observed when looking at rare and novel variants located in regions linked to the gene promoter by Hi-C (S15 Fig). The same increasing trend for DHSs correlated with promoters is observed for transcripts with different levels of AI (Fig 3C). However, the observed trend is not as strong in all transcripts (S16 Fig). Coding rare and novel variants especially in transcripts with homozygous eSNP also appear to have an impact on gene

expression, as they are as enriched in coding regions compared to common variants (Fig 3B). The effect is almost as strong as the one observed for non-coding variants located at DHSs highly correlated with the promoter (Pearson's r^2>0.9) (Fig 3B). Also, a similar trend of significant increased AI is observed for coding variants in transcripts with homozygous eSNP (Fig 3C, linear regression slope=0.227, p-value=0.018).

Having the advantage of higher power using this larger cohort, we observed that the more rare or novel variants there are within the vicinity of the transcribed region of a gene, the higher the likelihood is that the transcripts will display AI (Fig 4A), which is not observed for common variants. Finally, we looked at the enrichment of rare or common variants around the TSS of transcripts with homozygous eSNP and observed a higher enrichment at +/- 50kb from the TSS for rare variants compared to common variants (Fig 4B).

Taken together we have shown that, rare and novel variants identified in human immune cells using the Immune-genetics sequencing capture panel are enriched in DHSs that are highly correlated to the promoters of transcripts and in the coding regions of highly differentially expressed transcripts, for which the top associated SNP is homozygous. We also observed enrichment of rare and novel variants in the vicinity of the TSS regions, and the more rare or novel variants there are, the stronger is the allelic imbalance of the gene expression.

2.5 Discussion

In this study, we used existing DHS mapping data to build a custom capture panel designed to enable efficient re-sequencing of key immune cell regulatory regions. Our "Immune-genetics sequencing" panel provides the means to comprehensively assess both coding and non-coding variation that could be implicated in the development of immune and inflammatory diseases. Because the method is based on sequencing rather than genotyping it allows direct cost effective assessment of both rare and common variation without any reliance on LD or the need for imputation. We have shown that with high sequencing coverage we are able to study novel non-coding variants in a confident way, which cannot be realized using whole exome sequencing, or would be prohibitively expensive using whole genome sequencing (WGS). The targeted regions included in the Immune-genetics sequencing panel overlap with GWAS hits in immune and inflammatory diseases and eQTLs of immune cells.

An inevitable drawback of the Immune-genetics sequencing design is its inability to capture variants of relevance to the disease of interest that map outside the targeted regions. This limitation is illustrated by disease-associated SNPs that are not included in the panel. Since the Immune-genetics sequencing panel was based exclusively on DHSs seen in immune cells, these missing associated SNPs could reflect regulatory effects that are associated with non-immune cell based aspects of the disease [204], e.g. gastrointestinal tract DHSs in ulcerative colitis. The fact that our panel captures the majority of the known immune and chronic inflammatory disease associated SNPs indicates that it will have broad utility across multiple immune related diseases.

Until now targeted capture methods have focused almost exclusively on the coding regions of the genome [227], which means that the effects of rare non-coding variants have largely been ignored in the analysis of complex traits. In our exploration of the approach we found that the

non-coding rare and novel variants identified by Immune-genetics sequencing frequently modify transcription factor binding motifs and show higher levels of selective constraint than are seen in included common sequence variants. This difference is expected based on evolutionary and population genetics principles, with common variants expected to be more neutral than rare ones [29].

A further novel aspect of the Immune-genetics sequencing approach is its inherent ability to utilise ASE information to interrogate the functional impact of sequence variants on gene expression. The greater power of ASE allowed us to observe functional effects using a lower sample size of unrelated subjects than traditional eQTL analysis [41]. In total the rare and novel variants identified by Immune-genetics sequencing explained 14% of the residual allelic imbalance in expression observed amongst individuals homozygous for common variants know to influence ASE, indicating that rare and novel variants likely account for at least part of the AI observed in the transcripts from individuals heterozygous for common eSNPs. Comparing non-coding variants in DHSs to variants in coding exons, the coding variants appeared to have a stronger effect on gene expression. However, the opposite situation was observed for variants located in DHSs that are correlated with gene promoters, where the effect of the non-coding variants was larger than those of coding ones. Rare and novel variants with substantial effects on AI in particular genes may contribute to certain disease phenotypes. In contrast to previous studies, we did not limit our exploration to extreme phenotypes, but instead we investigated the whole spectrum of AI. In doing so, we observed that the effect of rare and novel variants on gene expression does not appear to be limited to extreme differences in allelic expression, but may also affect genes with moderate AI.

One further limitation of the study may be that not all transcripts for which the allelic expression is skewed were accounted for by rare variants identified by Immune-genetics sequencing. Some variants exerting long range or trans effects will inevitably have been missed by not performing WGS. Nevertheless, as opposed to earlier studies [48, 49], we expand the exploration of rare variant effects to distal regulatory sites with correlated activity with gene promoter. While distal sites show enrichment, the strongest effect of rare and novel variants is found around the TSS of genes displaying AI. This observation indicates that variants can be clustered to perform collapsing association test for complex traits, which will permit the identification of rare and novel trait-associated variants and to easy linking of the variants to a specific gene.

2.6 Conclusion

In this study, we show that targeted re-sequencing of cell specific active regulatory regions can be an efficient means to identify functionally relevant variation that is considerably more cost effective than WGS. Immune-genetics sequencing provides an efficient means to identify rare and novel, coding and non-coding variation of relevance in complex traits involving the immune system and to study the impact of rare and novel non-coding regulatory variants on other epigenetic traits.

Abbreviations

AI: Allelic imbalance; ASE: Allele-specific expression; CADD: Combined annotation dependent depletion; DHS: DNase I hypersensitive sites; DNase I: Deoxyribonuclease I; DNase-seq: DNase I sequencing; eQTL: Expression quantitative trait loci; eSNP: SNPs altering expression; FIMO: Finding Individual Motif Occurrence; GERP: Genomic Evolutionary Rate Profiling; GWAS: Genome-wide association study; HLA: Human leucocyte antigen; MAF: Minor allele frequency; MHC: Major histocompatibility complex; NGS: Next-generation sequencing; RNA-seq: RNA sequencing; SNP: Single nucleotide polymorphism; TSS: Transcription start site; WES: Whole-exome sequencing; WGS: Whole-genome sequencing

2.7 Acknowledgements

This work was supported by grants from the Canadian Institute of Health Research (CIHR), the UK Medical Research Council (G1100125), the Swedish Research Council (DO283001) and Knut and Alice Wallenberg Foundation (KAW). We also acknowledge the use of subjects from the Cambridge BioResource and the support of the Cambridge NIHR Biomedical Research Centre. AM was supported by the Fond de Recherche Santé Québec Doctoral training award. TP and CL hold a Canada Research Chair.

Declarations

Supporting information is available online

Availability of data and materials

All data are available through EGA (https://www.ebi.ac.uk/ega/home) under the study "Immune-genetics sequencing" (DAC: EGAC00001000409, Policy: EGAP00001000396, Study: EGAS00001001564).

Authors' contributions

TP, GB and ML conceived and supervised the study. AM, TP and TK drafted the manuscript. AM, TK, BG, MC and LL analyzed the data. MB, KT, JS, JC, A-CS, LR, SJS provided samples and materials. All authors reviewed and approved the final manuscript.

Competing interest

Authors declare no competing interests

Consent for publication

Not applicable

Ethics approval and consent to participate

Written and informed consent was obtained for each participant during enrollment and was approved by each participating sites' regional ethical review board.

2.8 Figures and Tables

Table 1. Sequencing statistics of the samples sequenced with Immune-genetics sequencing

	Mean target coverage	Bases on target (%)[1]	Target region without coverage (%)[2]	Target bases with >=10x coverage (%)[3]	Level of multiplexing	Sequencing platform
Sweden Uppsala Bioresource samples (n=30)	52X	88	1.9	83	2X (3 samples) 5X (27 samples)	HiSeq2500 (2X samples) HiSeq2000 (5X samples)

Alignment to the human hg19 reference genome, and variant calling (HaplotypeCaller) to identify all SNPs were performed. Shows average values across samples. [1] On and near bait bases/good quality bases aligned (according to Picards metrics). [2]The percentage of target region that did not reach 2x coverage over any base.[3] The percentage of all target bases achieving 10X or higher coverage. We considered a variant to be true at >=10 depth.

Table 2. General characteristics of the common, rare and novel single nucleotide variations (SNVs)

Total number (average per sample)		All	Common	Rare	Novel[1]
All (Immune-genetics sequencing)		351088 (90594)	275042 (83839)	50004 (5318)	26042 (1437)
Coding[2]	All	60946 (15169)	45545 (1818)	12,452 (1166)	2949 (185)
	Non-synonymous[3]	30967 (7174)	21807 (6403)	7405 (669)	1755 (102)
	Synonymous[3]	29214 (7770)	23434 (7305)	4785 (395)	995 (71)
	Stop-gained[3]	395 (71)	202 (56)	135 (13)	58 (2)
Exome[4]		120245 (30682)	91818 (27916)	21497 (2280)	6,930 (486)
Non-coding[5]		290142 (75424)	229497 (70020)	37552 (4152)	23093 (1251)
All DHS[6]		195182 (51559)	154154 (48056)	24571 (2677)	16457 (826)

Total number of variants and the average number of variants per sample that were included in the Immune-genetics sequencing design. [1]Novel variants are defined as not identified in the 1000 Genomes Project nor included in dbSNP141. [2]Coding variants are those located in the exons of the RefSeq coding sequence. [3]Synonymous, non-synonymous and stop-gained variants were annotated using SNPeff and the hg19 version of the genome. [4]The Exome is based on the Roche SeqCap EZ exome v3.0. [5] Non-coding variants are those not in the RefSeq coding sequence. [6]The All DHSs category combines all DHSs from the selected 12 cell types and could partly overlap with the Exome. Cut-offs used for the quality control of the variants are read depth≥10, genotyping quality (gq)≥70, mapping quality (MQ) ≥50, and proportion of the reference allele between 10-90%.

Figure 1

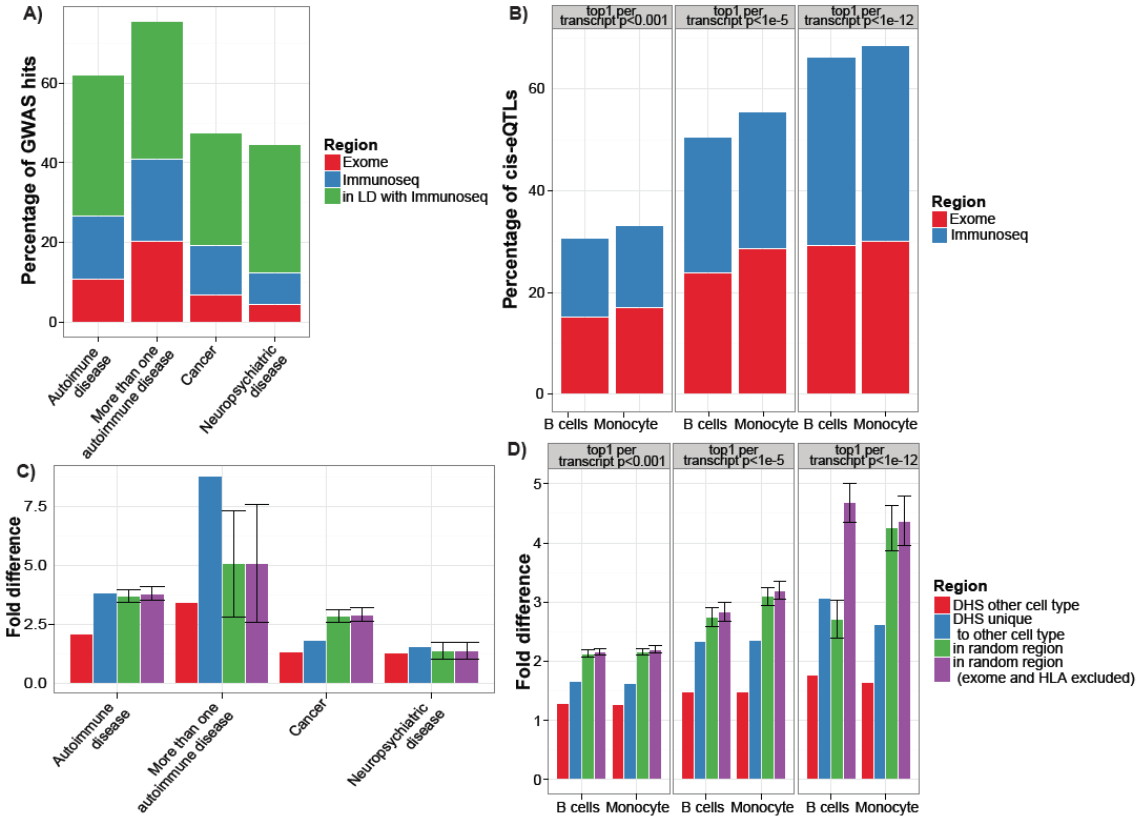


Fig 1. Benchmarking the Immune-genetics sequencing capture panel by known disease associated sites and regulatory variants. (**A**) Autosomal GWAS hits associated to more than one autoimmune or chronic inflammatory disease, for neuropsychiatric diseases and for cancer included in the Immune-genetics sequencing custom capture panel. (Cut-off of $1x10^{-8}$ was used to select GWAS hits to analyze, SNPs in LD selected based on r^2>0.9, HLA (human leucocyte antigen) hits and region as well as chromosome X SNPs were excluded from the analyses). SNP in LD = GWAS hits that have a SNP in LD in the Immune-genetics sequencing custom capture panel. (**B**) cis-eQTLs from monocytes (CD14+) and B Cells (CD19+) (considered has haplotype block, r^2>0.9) included in the Immune-genetics sequencing panel. Cut-off of p<1e-3 or p<1e-5, and p<1e-12 after 1000 permutations (1000= number of SNPs tested per probe) and top 1 eQTLs per transcript were kept for analysis (HLA hits and region as well as chromosome X hits were excluded in the analyses). (**C**) Enrichment of GWAS hits (same as in A) and proximal SNPs (LD r^2>0.9) that fall in DHSs selected for immune cell types compared to DHSs selected from other tissues (either all or non-overlapping ones) and regions randomly selected (1000 times) from the whole genome (either the full genome or only non-coding regions excluding HLA). Significance was calculated using Fisher´s exact test. Enrichment is significant (p<0.001) for all GWAS hits except for Neuropsychiatric hits. (**D**) Enrichment of eQTLs (same as in B) and proximal SNPs (LD r^2>0.9) positioned at DHSs selected for immune cell types compared to DHSs selected from other tissues (either all or non-overlapping ones) and regions randomly selected (1000 times) from the whole genome (either entire genome or only the non-coding part excluding the HLA region). All enrichments shown are significant (p<0.001). All p-values were calculated using Fisher´s exact test.

Figure 2

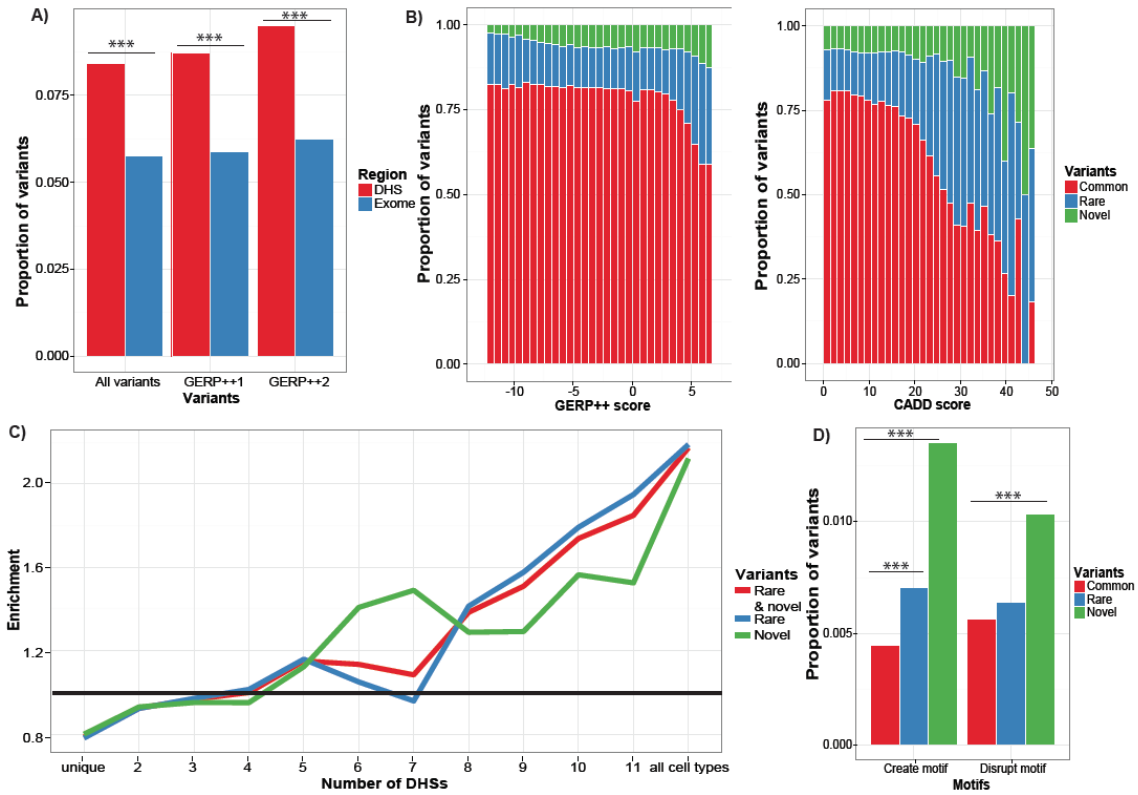


Fig 2. Discovery and functional potential of rare and novel variants using Immune-genetics sequencing. (**A**) Proportion of novel variants (all, Genomic Evolutionary Rate Profiling (GERP++)>=1 and GERP++>=2) identified in DHS (red) compared to the exome (blue). (**B**) Distribution of proportion of common (red), rare (blue) and novel (green) variants according to GERP++ score and Combined annotation dependent depletion (CADD) score. (**C**) Fold enrichment of rare (blue), novel (green) or rare and novel combined (red) variants compared to common variants found at shared or cell-type specific DHSs. Linear regression slope: rare =0.119 p-value= 1.35e-05, novel= 0.093 p-value= 5.81e-05, rare and novel=0.113 p-value=2.41e-06. (**D**) Proportion of common (red), rare (blue) and novel (green) variants localized at a DHS that either disrupt or create a transcription-factor binding motif. P-values are calculated using Fisher's exact test (***p<0.001).

Figure 3

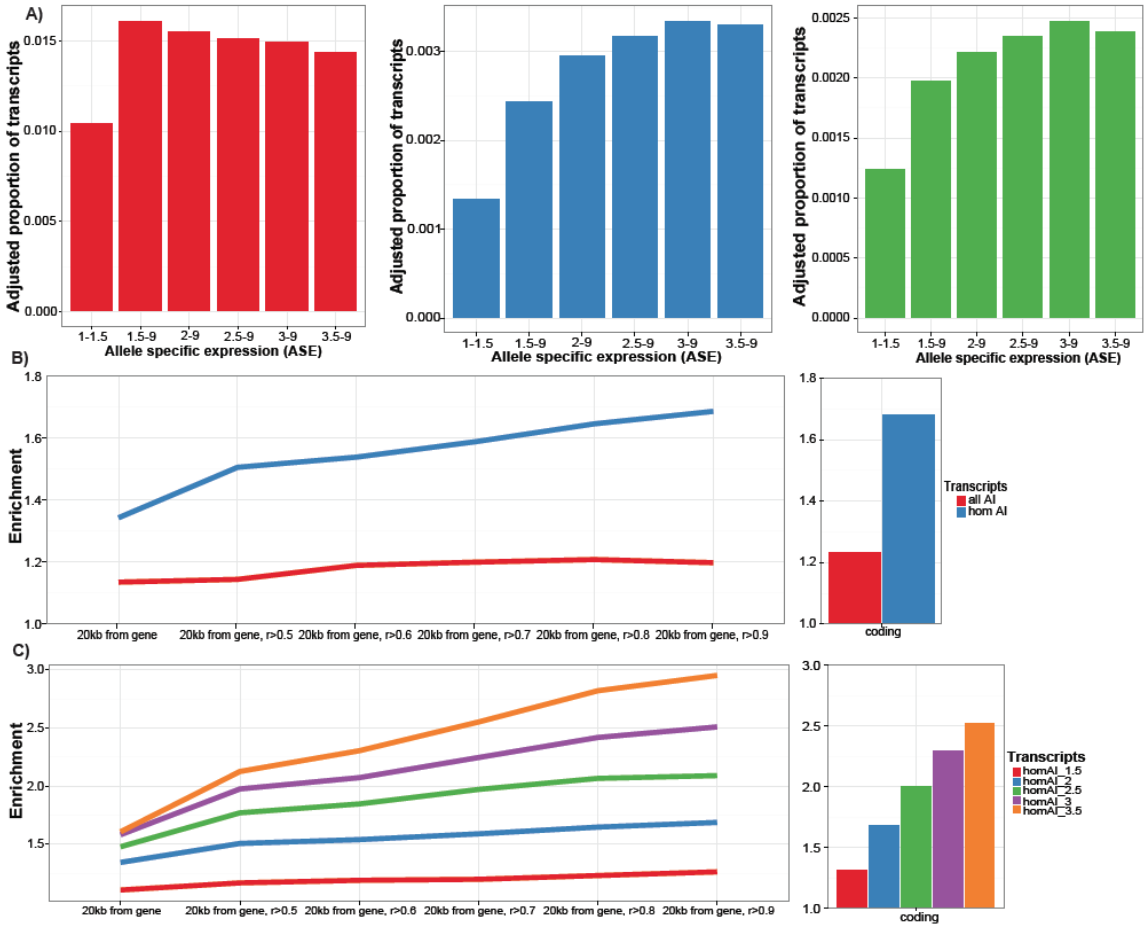


Fig 3. The impact of rare and novel noncoding variants on gene expression. (**A**) Using the replication set, we looked at the adjusted proportion of transcripts with common (red), rare (blue) or novel (green) noncoding variants in the vicinity (+/-20kb) of a gene based on different allelic imbalance: 1.5 to 9, 2 to 9, 2.5 to 9, 3 to 9 and 3.5 to 9-fold difference. Adjustment was based on average number of SNPs used to calculate ASE at each ASE levels. (**B**) Enrichment of proportion of transcripts showing allelic imbalance (AI) with rare or novel variants in the vicinity of the gene compared to AI transcripts with common variants in vicinity of a gene. We looked at coding (histogram) vs noncoding variants as well as noncoding variants in DHS regions correlated with the promoters (Pearson correlation r>0.5 to 0.9). In red are all transcripts where allelic imbalance was measured (allAI) and in blue are the transcripts for which the top associated SNP is homozygous in the sample (homAI). Linear regression slope for allAI= 0.015 (p-value=0.0196) and homAI=0.063 (p-value= 0.0024). Allelic imbalance genes are considered as >=2 fold between the alleles and equally expressed genes are <=1.5 fold. (**C**) Fold difference between proportions of AI transcripts with rare or novel variants in the vicinity compared to AI transcripts with common variants in the vicinity. Only including transcripts for which the top associated SNP is homozygous (homAI). We looked at coding (histogram) vs noncoding variants around the genes (+/-20kb from gene) and in DHS regions correlated with the promoters (Pearson correlation r>0.5 to 0.9). We compare different levels of allelically imbalanced transcripts from 1.5-fold to 3.5. all AI: AI transcripts comparing all transcripts for which ASE

was measured and homAI: transcripts for which the top associated SNP that drives the association across samples is homozygous.

Figure 4

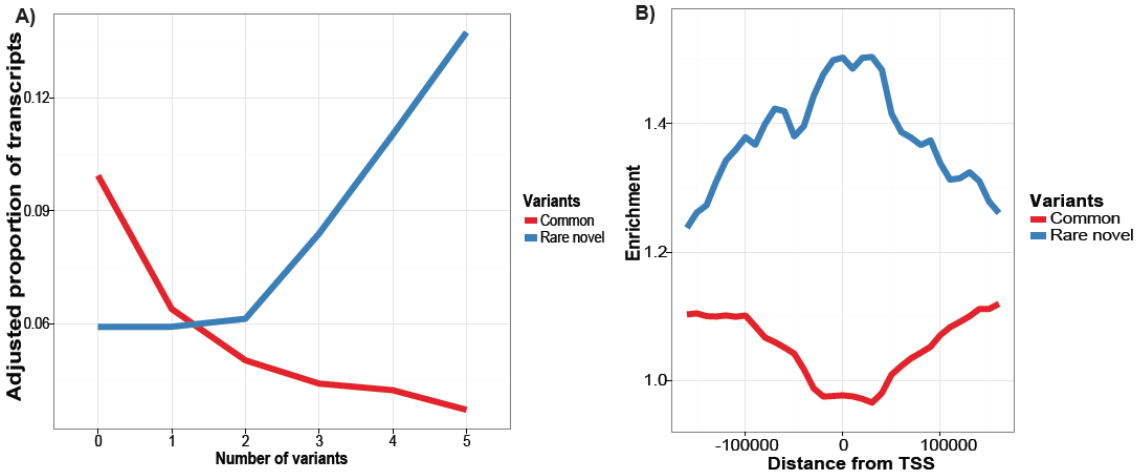


Fig 4. The number and location of rare and novel noncoding variants have an impact on gene. (**A**) Adjusted proportion of AI transcripts that contain 1 or more noncoding common (red) or rare and novel (blue) variants in transcripts vicinity (+/-20kb from gene). Adjustment was based on average number of SNPs used to calculate ASE at each ASE levels. (**B**) Fold enrichment of common (red) or rare and novel (blue) variants in AI vs all transcripts measuring their distance from transcription start sites (TSS). Transcripts with $p<0.05$ were used. Sliding window of 80kb every 10kb was used.

2.9 Supporting information

Table S1. Cell type selected to target regulatory regions in immune cells.

Cell types	Number of bins	Final number of bins	Number of samples	Accession number (GEO)
CD3+	250,000	259,321	4	GSM665837, GSM701488, GSM701516, GSM774201
CD3+ cord blood	150,000	97,243	2	GSM701525, GSM701526
CD4+	200,000	196,585	8	GSM665812, GSM665839, GSM701489, GSM701491, GSM701539, GSM817166, wgEncodeUwDnaseCd4naivewb11970640AlnRep1, wgEncodeUwDnaseCd4naivewb78495824AlnRep1
CD8+	250,000	229,426	5	GSM665813, GSM665838, GSM701499, GSM701540, GSM817160
CD14+	250,000	271,497	4	GSM701503, GSM701541, wgEncodeUwDnaseMonocd14ro1746AlnRep1V2, wgEncodeUwDnaseMonocd14ro1746AlnRep2
CD19+	250,000	240,941	3	GSM701492, GSM701493, GSM701507
CD20+	225,000	200,000	1	GSM701500
CD34+	250,000	242,494	14	GSM493384, GSM493386, GSM493387, GSM530652, GSM530657, GSM530658, GSM530659, GSM530660, GSM530663, GSM530664, GSM595914, GSM595917, GSM595918, GSM595919
CD56+	200,000	183,311	3	GSM665820, GSM665836, GSM701508
Th1	100,000	64,012	2	wgEncodeOpenChromDnaseAdultcd4th1AlnRep1, wgEncodeOpenChromDnaseAdultcd4th1AlnRep2
Th2	300,000	291,859	1	GSM736502
Th17	100,000	100,000	1	wgEncodeUwDnaseTh17AlnRep1 (to verify)

Table S2. Cell types selected to target regulatory regions in other cell types not related to immune function.

Cell types/Tissue	Number of bins	Final number of bins	Number of samples	Accession number (GEO)
Fetal Lung	250,000	231,527	11	GSM530662, GSM595915, GSM595916, GSM595921, GSM595924, GSM595925, GSM595927, GSM595929, GSM595930, GSM665805, GSM665806
Fetal Kidney	200,000	193,630	6	GSM493385, GSM774221, GSM817159, GSM878666, GSM1024608, GSM1027329
Fetal Brain	200,000	173,884	9	GSM530651, GSM595913, GSM595920, GSM595922, GSM595923, GSM595926, GSM595928, GSM665804, GSM1027328
Fetal Small intestine	250,000	225,102	11	GSM665825, GSM665835, GSM701487, GSM701496, GSM701504, GSM701530, GSM774205, GSM774210, GSM774216, GSM817161, GSM817187
Fetal Large intestine	250,000	233,707	9	GSM701490, GSM701495, GSM701531, GSM774213, GSM774214, GSM774217, GSM774220, GSM817162, GSM817188
Fetal Renal cortex	250,000	245,861	10	GSM701494, GSM701502, GSM701529, GSM701532, GSM817176, GSM878629, GSM878667, GSM1027314, GSM1027316, GSM1027323
Fetal Stomach	250,000	232,109	13	GSM701498, GSM701521, GSM701528, GSM701538, GSM774202, GSM774212, GSM817173, GSM817199, GSM878660, GSM878665, GSM1024606, GSM1027318, GSM1027331
Fetal Arm muscle	250,000	241,617	15	GSM701506, GSM701535, GSM774223, GSM774239, GSM817178, GSM817184, GSM817214, GSM817216, GSM878610, GSM878618, GSM878619, GSM878620, GSM878625, GSM878638, GSM1024605
Fetal Placenta	250,000	231,726	5	GSM774215, GSM774219, GSM817219, GSM878659, GSM1027343
Fetal Adrenal gland	200,000	179,763	5	GSM817165, GSM817167, GSM878658, GSM1027310, GSM1027311
Fetal Testis	250,000	198,545	2	GSM878617, GSM1027319
Fetal Ovary	100,000	100,000	1	GSM1027306

Table S3. Summary of shared common, rare and novel variants in selected DHS regions of different immune cells

Common/ Rare/ novel	CD3+	CD3+ cord blood	CD4+	CD8+	CD14+	CD19+	CD20+	CD34+	CD56+	Th1	Th2	Th17
CD3+	54,570/ 9,834/ 5,774											
CD3+ cord blood	21,628/ 3,880/ 2,469	21,730/ 3,909/ 2,481										
CD4+	40,663/ 7,288/ 4,416	21,421/ 3,849/ 2,436	42,354/ 7,520/ 4,561									
CD8+	45,291/ 7,995/ 4,885	21,385/ 3,827/ 2,442	38,704/ 6,840/ 4,193	49,086/ 8,602/ 5,239								
CD14+	26,697/ 4,973/ 3,062	18,040/ 3,328/ 2,083	25,084/ 4,585/ 2,819	25,886/ 4,746/ 2,948	59,581/ 9,829/ 5,819							
CD19+	35,476/ 6,364/ 3,992	21,245/ 3,822/ 2,428	32,055/ 5,731/ 3,596	34,235/ 6,110/ 3,862	27,704/ 5,087/ 3,150	50,942/ 8,773/ 5,407						
CD20+	20,035/ 3,654/ 2,256	14,707/ 2,657/ 1,677	18,817/ 3,448/ 2,127	19,488/ 3,519/ 2,172	18,709/ 3,385/ 2,089	19,000/ 3,441/ 2,143	43,910/ 7,047/ 4,288					
CD34+	31,128/ 5,728/ 3,548	19,886/ 3,586/ 2,295	28,223/ 5,149/ 3,138	29,834/ 5,437/ 3,408	31,088/ 5,544/ 3,461	29,577/ 5,366/ 3,402	20,419/ 3,654/ 2,276	53,076/ 9,165/ 5,552				
CD56+	35,904/ 6,434/ 3,980	20,571/ 3,684/ 2,360	32,535/ 5,783/ 3,633	35,943/ 6,387/ 3,975	24,905/ 4,550/ 2,787	29,661/ 5,290/ 3,371	18,076/ 3,343/ 2,064	27,573/ 5,033/ 3,138	40,046/ 7,156/ 4,359			
Th1	12,015/ 2,171/ 1,304	9,770/ 1,714/ 1,082	11,479/ 2,062/ 1,265	11,846/ 2,119/ 1,280	10,734/ 1,930/ 1,179	11,060/ 1,997/ 1,224	9,063/ 1,638/ 990	11,365/ 2,044/ 1,252	11,181/ 2,001/ 1,232	13,025/ 2,412/ 1,414		
Th2	36,811/	18,595/	32,002/	34,935/	24,458/	30,582/	18,023/	27,643/	29,154/	11,768/	59,458/	

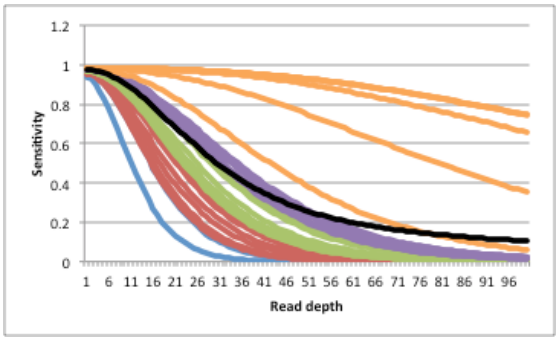
	6,280/ 3,903	3,285/ 2,132	5,456/ 3,451	5,876/ 3,742	4,337/ 2,716	5,220/ 2,392	3,166/ 1,964	4,895/ 3,096	5,018/ 3,182	2,043/ 1,250	9,775/ 5,707	
Th17	31,088/ 3,260/ 1,979	11,617/ 2,254/ 1,443	15,581/ 3,010/ 1,847	16,234/ 3,076/ 1,894	13,322/ 2,665/ 1,589	13,051/ 2,581/ 1,596	9,967/ 1,949/ 1,207	14,468/ 2,808/ 1,713	15,758/ 2,955/ 1,867	7,574/ 1,417/ 907	17,141/ 3,186/ 1,922	22,026/ 4,146/ 2,366

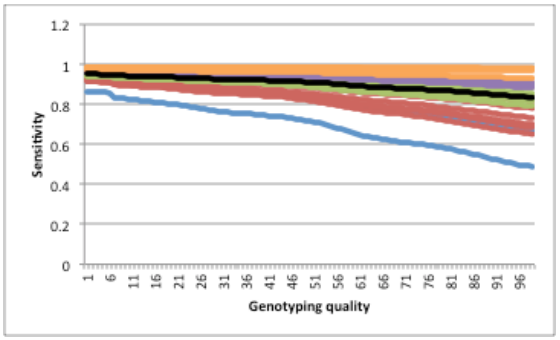
Table S4. Sequencing statistics of the Cambridge Multiple sclerosis samples with Immune-genetics sequencing.

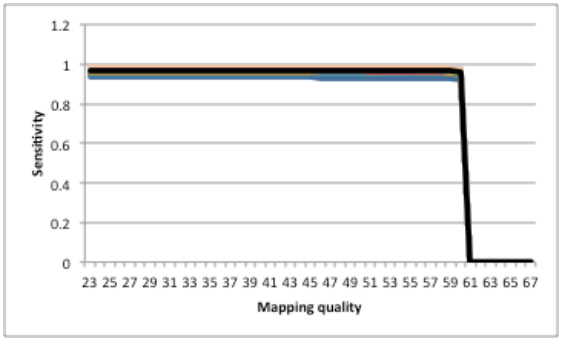
	Mean target coverage	Bases on target (%)[1]	Target region without coverage (%)[2]	Target bases with >=10x coverage (%)[3]	Sequencing platform	Level of multiplexing
Cambridge Multiple Sclerosis cohort and healthy controls (n=180)	31X	68.76	1.9	73	HiSeq2000	6X

Alignment to the human hg19 reference genome, and variant calling (HaplotypeCaller) to identify all SNPs were performed. Shows average values across samples. [1] On and near bait bases/good quality bases aligned (according to Picards metrics). [2]The percentage of target region that did not reach 2x coverage over any base.[3] The percentage of all target bases achieving 10X or higher coverage. We considered a variant to be true at >=10 depth.

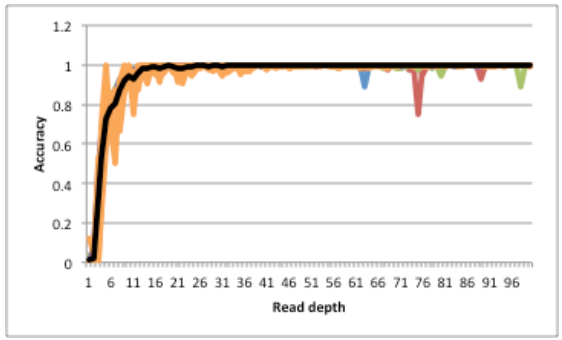
A)

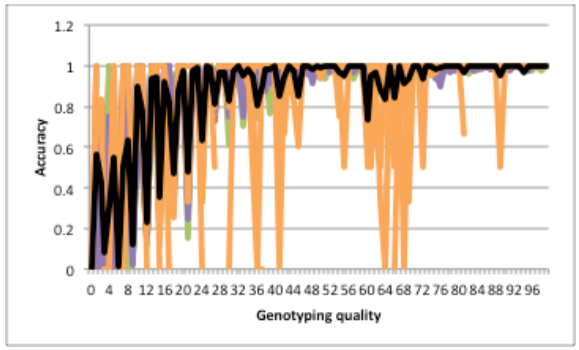

Sensitivity
Read depth

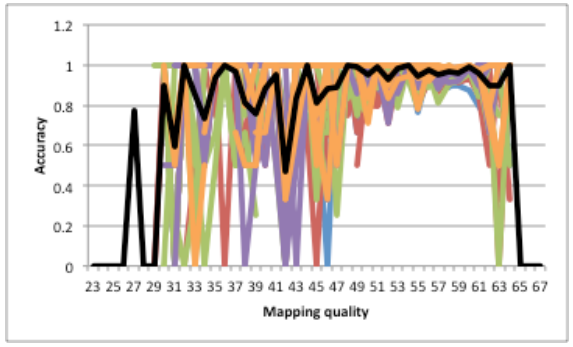

Sensitivity
Genotyping quality


Sensitivity
Mapping quality

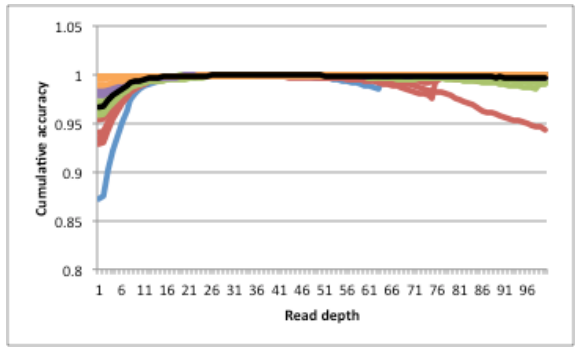
B)


Accuracy
Read depth


Accuracy
Genotyping quality


Accuracy
Mapping quality

C)


Cumulative accuracy
Read depth

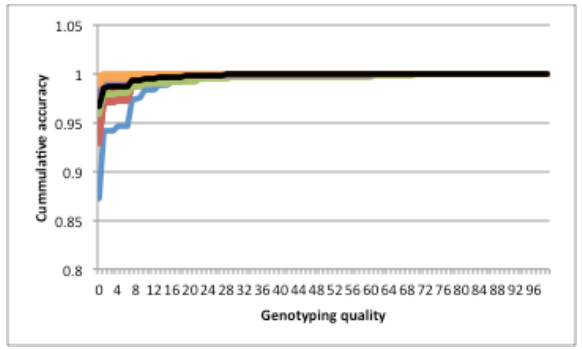

Cummulative accuracy
Genotyping quality

Figure S1. Variants Quality control

Comparing sequencing data to genotyping data (Human Omni2.5 BeadChip) considering only heterozygous SNPs for the discovery set. In black is the average, samples are coloured according to mean coverage: blue=15-20, red=20-30, green=30-40, purple= 40-50, orange>=50. **A**) Sensitivity: how many of the genotyped SNPs are called in the Immune-genetics sequencing at increasing read depth/genotyping quality/mapping quality. **B)** Accuracy: Of the variant captured at each read depth/genotyping quality/mapping quality, how many of them are accurate. **C)** Cumulative accuracy: Of the variant captured at each read depth/genotyping quality and over, how many of them are accurate.

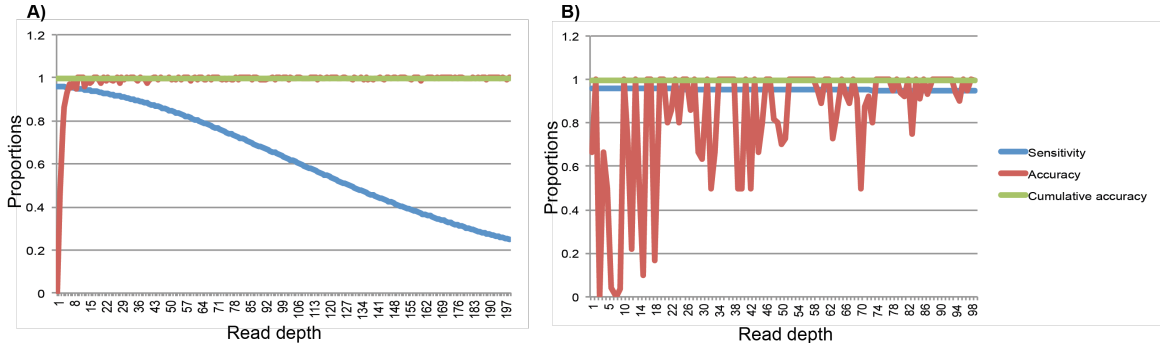


Figure S2. Comparing sequencing data for NA18502 sample (complete genomics data and Immune-genetics sequencing) considering only heterozygous SNVs identified by complete genomics that fall into Immune-genetics sequencing custom capture panel. Sensitivity: how many of the genotyped SNP are we seeing in the Immune-genetics sequencing at increasing read depth/genotype quality (ex: over 10 read depth we capture 95% of the heterozygous variants). Accuracy: Of the variant captured at each depth (and over), how many of them are accurate. **A**) Read depth. **B**) Genotyping quality.

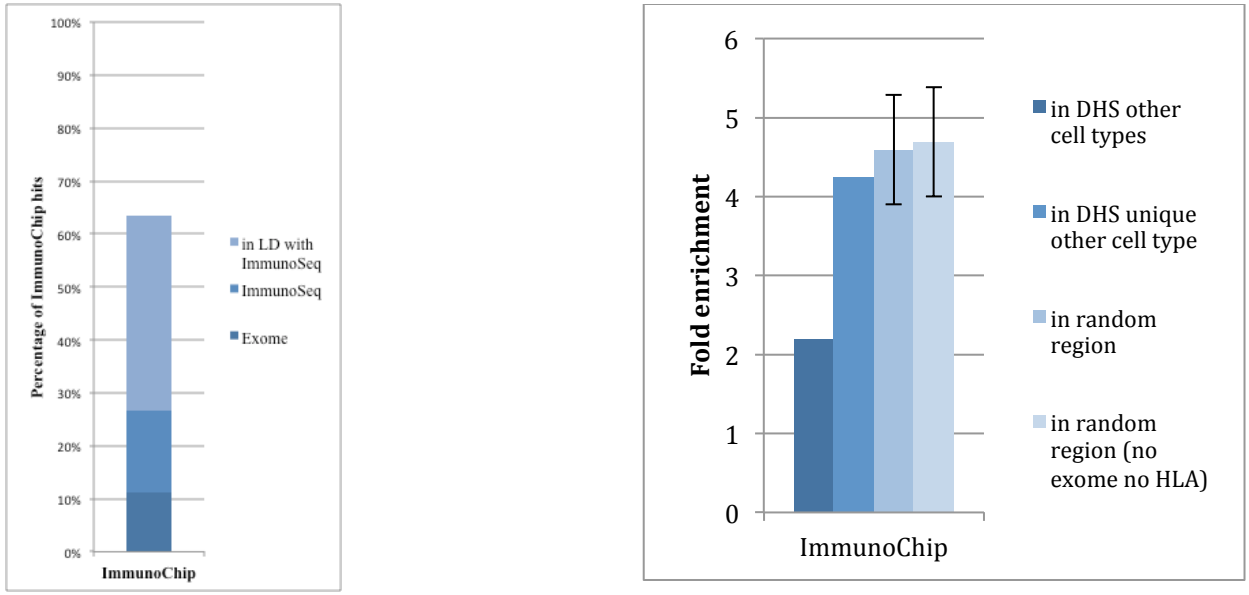


Figure S3. ImmunoChip hits that falls into Immune-genetics sequencing custom capture panel. A) ImmunoChip hits that falls in the Immune-genetics sequencing. capture panel. (Cut-off of $1x10^{-8}$ was used to select hits to analyze, SNPs in LD selected based on $r^2>0.9$, HLA hits and region as well as chromosome X SNPs were excluded from the analyses). SNP in LD = ImmunoChip hits that have a SNP in LD represented by Immune-genetics sequencing. **B**) Enrichment of hits (same as in A) and proximal SNPs (LD $r^2>0.9$) that fall in DHSs selected for immune cell types compared to DHSs selected from other tissues (either all or non-overlapping ones) and regions randomly selected (1000 times) from the whole genome (either entire genome or only non-coding excluding the HLA region). Significance was calculated using Fisher exact test.

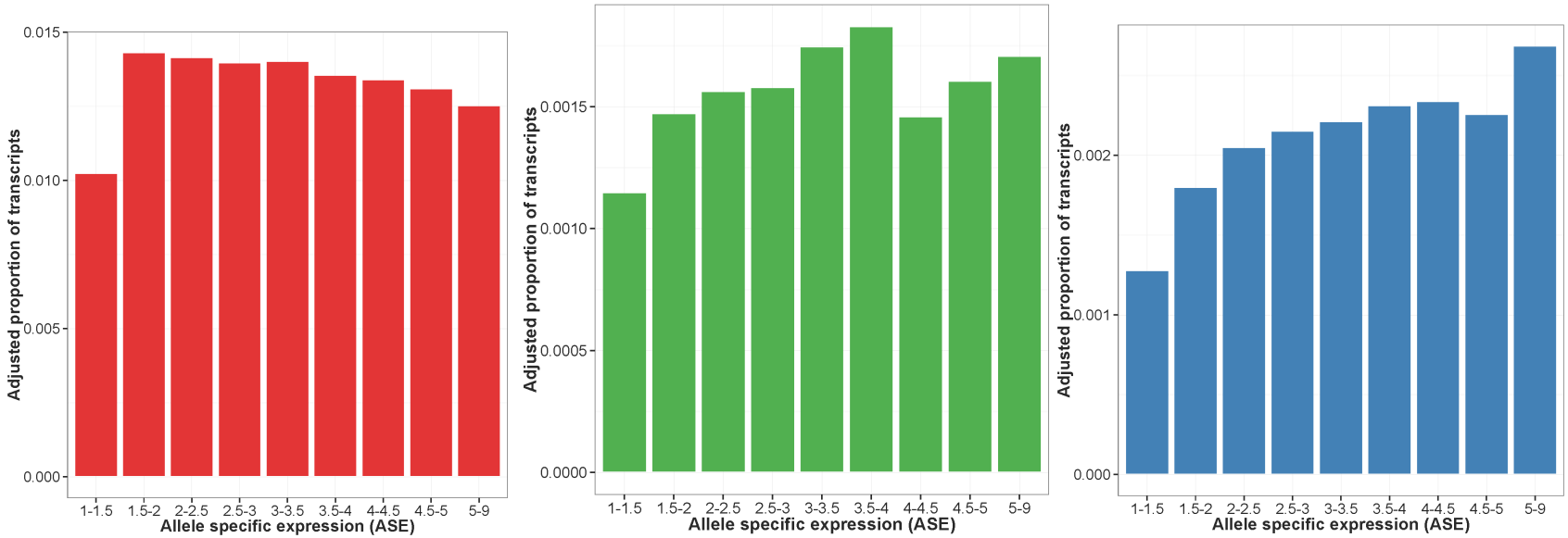


Figure S4. Discovery set distribution of allele specific expression (ASE). Distribution of proportion of ASE in transcripts with common (red), rare (blue) or novel (green) noncoding variants in vicinity (+/-20kb from gene) adjusted for average number of SNPs used to calculate ASE.

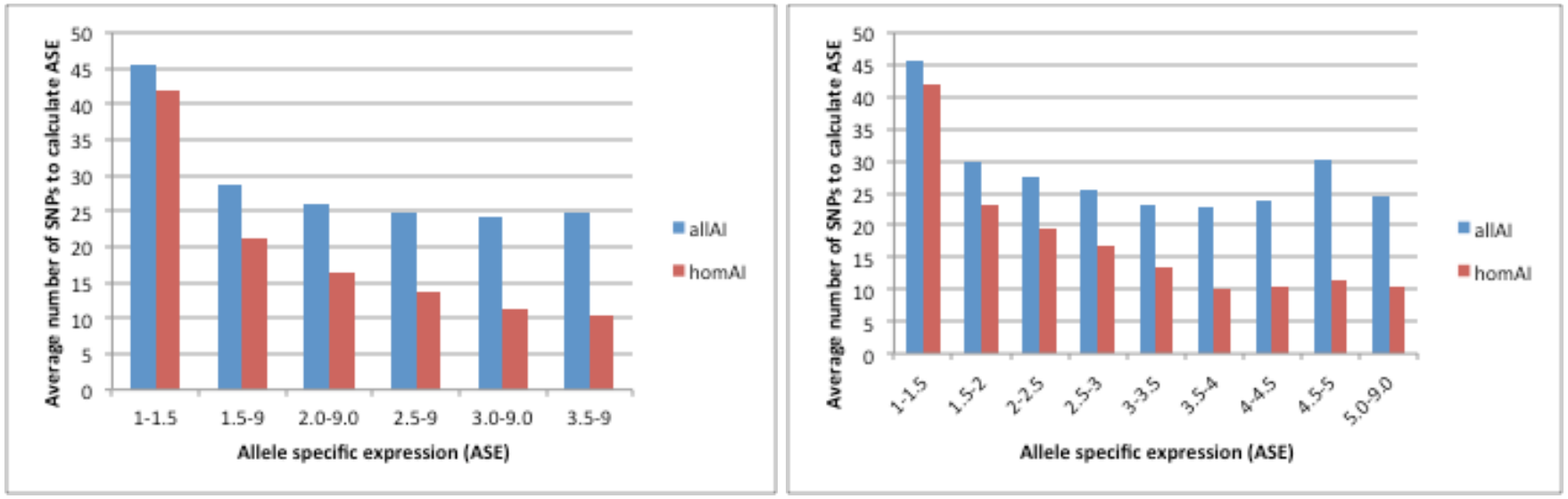


Figure S5. Average number of SNPs used to calculate allele specific expression (ASE) in discovery set samples. Comparing all transcripts for which ASE was measured (allAI) and transcripts for which the top associated SNP that drives the association across samples is homozygous (homAI).

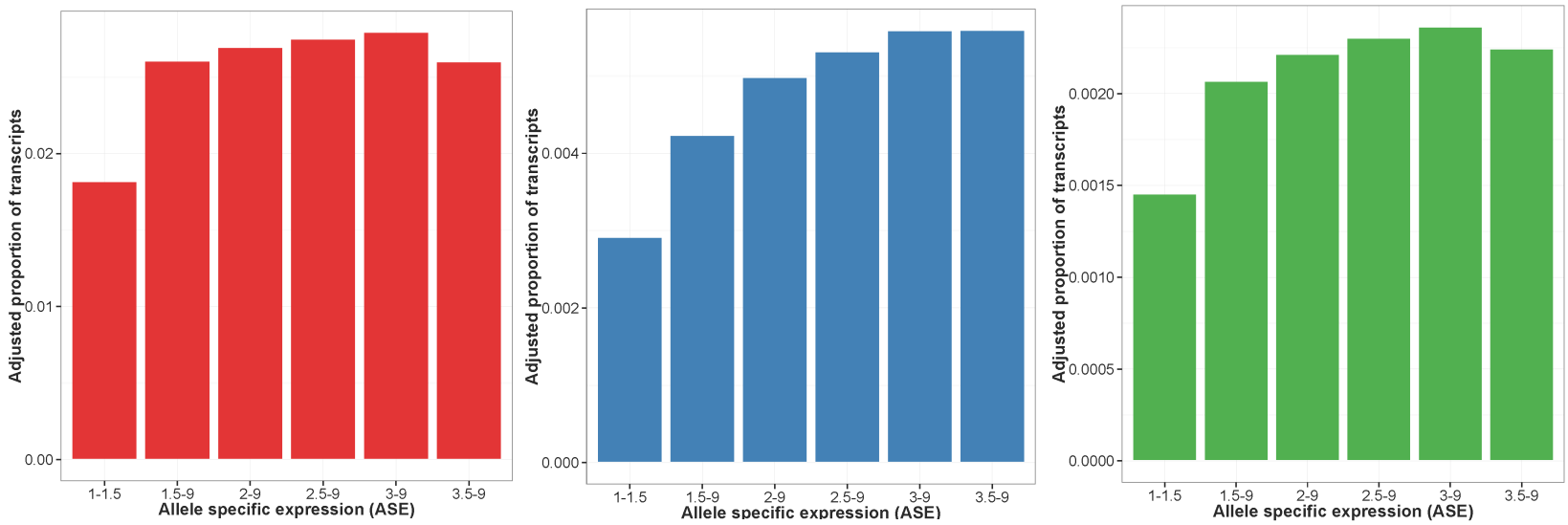


Figure S6. Adjusted proportion of transcripts with common (red), rare (blue) or novel (green) noncoding variants in the vicinity (+/-20kb) of a gene based on different allelic imbalance: 1.5 to 9, 2 to 9, 2.5 to 9, 3 to 9 and 3.5 to 9-fold difference in the discovery set. Adjustment was based on average number of SNPs used to calculate ASE at each ASE levels.

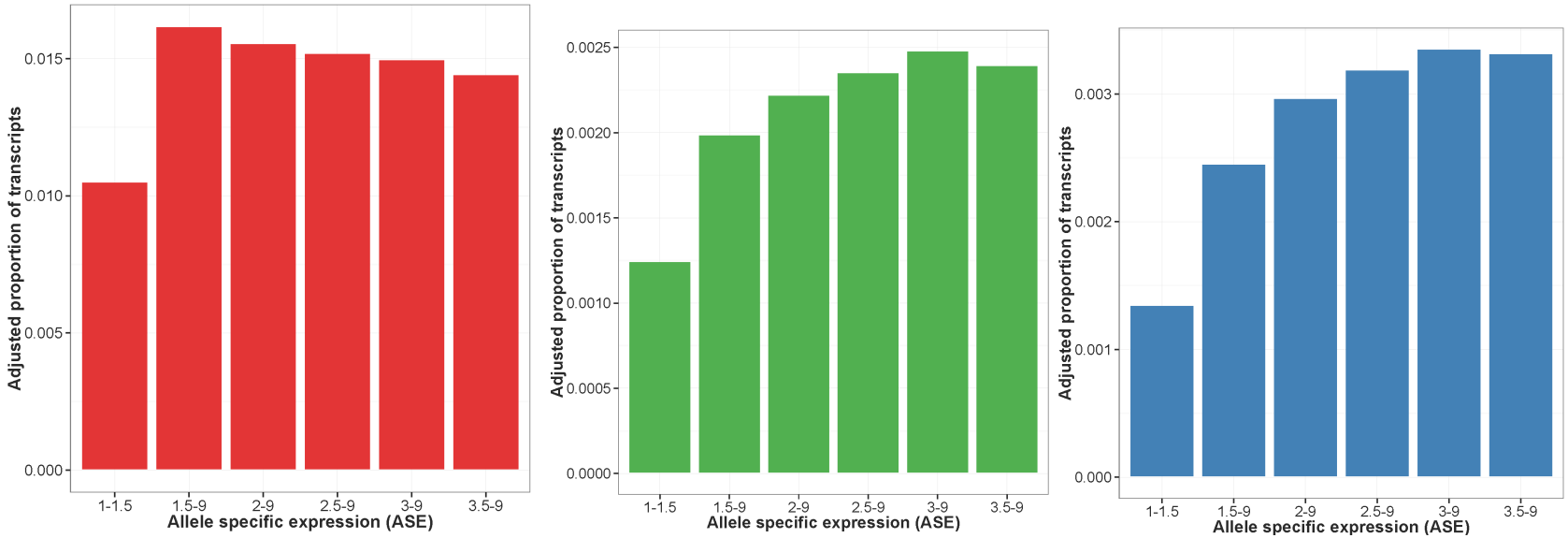


Figure S7. Discovery set distribution of Allelic imbalance (AI). Adjusted proportion of transcripts with common (red), rare (blue) or novel (green) noncoding variants in vicinity (+/-20kb from gene) based on different AI: 1.5 to 9, 2 to 9, 2.5 to 9, 3 to 9 and 3.5 to 9-fold difference. Including only transcripts for which the top associated SNP is homozygous in the sample (homAI).

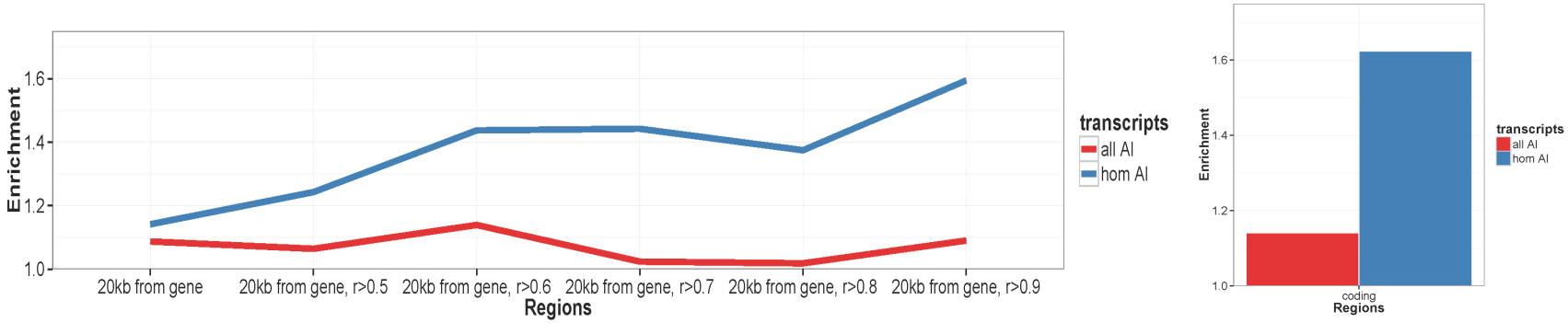


Figure S8. Enrichment of proportion of AI transcripts with rare or novel variants in vicinity of a gene compared to AI transcripts with common variants in vicinity of a gene in the discovery set. We looked at coding (histogram) vs noncoding variants as well noncoding variants in DHS region correlated with the promoter (Pearson correlation r>0.5 to 0.9). In red are all transcripts where allelic imbalance was measured (allAI) and in blue are the transcripts for which the top associated SNP is homozygous in the sample (homAI). Linear regression slope for homAI=0.076 (p-value= 0.018) and allAI= -0.007 (p-value= 0.591). Allelic imbalance genes are considered as >=2 fold between the alleles and equally expressed genes are <=1.5 fold.

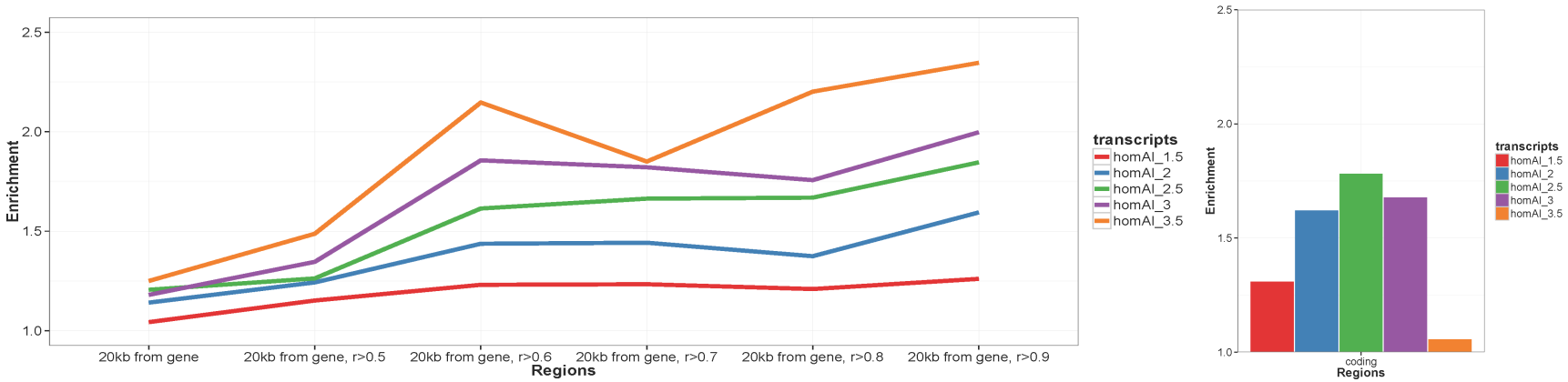


Figure S9. Fold difference between proportion of AI transcripts with rare or novel variants in vicinity compared to AI transcripts with common variants in vicinity in the discovery set. Only including transcripts for which the top associated SNP is homozygous (homAI). We looked at coding (histogram) vs noncoding variants around the genes (+/-20kb from gene) and in DHS regions correlated with the promoters (Pearson correlation r>0.5 to 0.9). We compare different level of allelically imbalanced transcripts from 1.5 fold to 3.5.

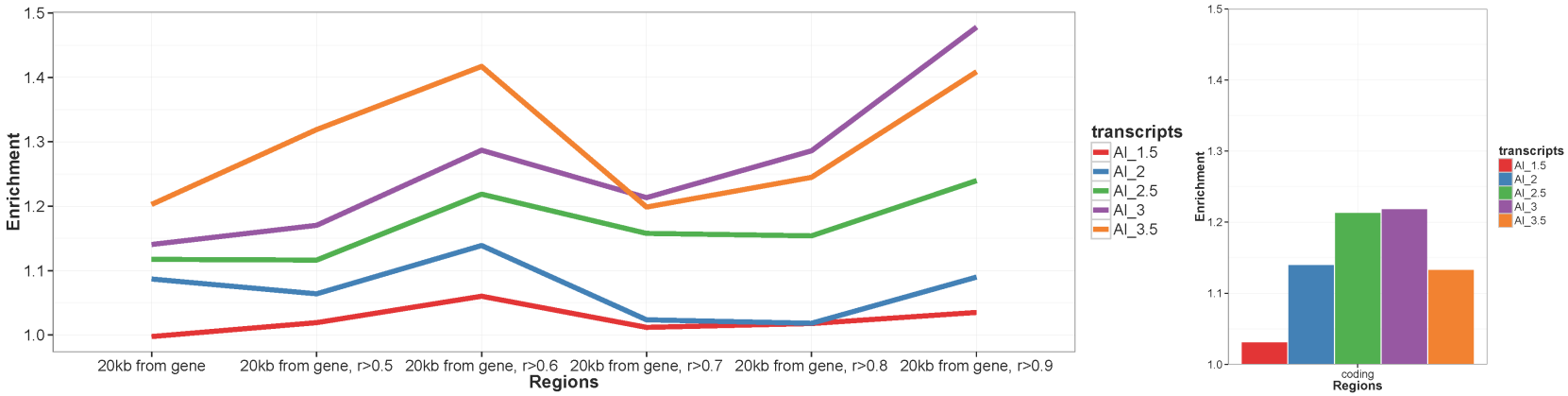


Figure S10. Enrichment between proportions of AI transcripts with rare or novel variants in vicinity compared to AI transcripts with common variants in vicinity in the discovery set. Including all transcripts (allAI). We looked at coding (histogram) vs noncoding variants around the genes (+/-20kb from gene) and in DHS regions correlated with the promoters (Pearson correlation r>0.5 to 0.9). We compare different levels of allelically imbalanced transcripts from 1.5 fold to 3.5.

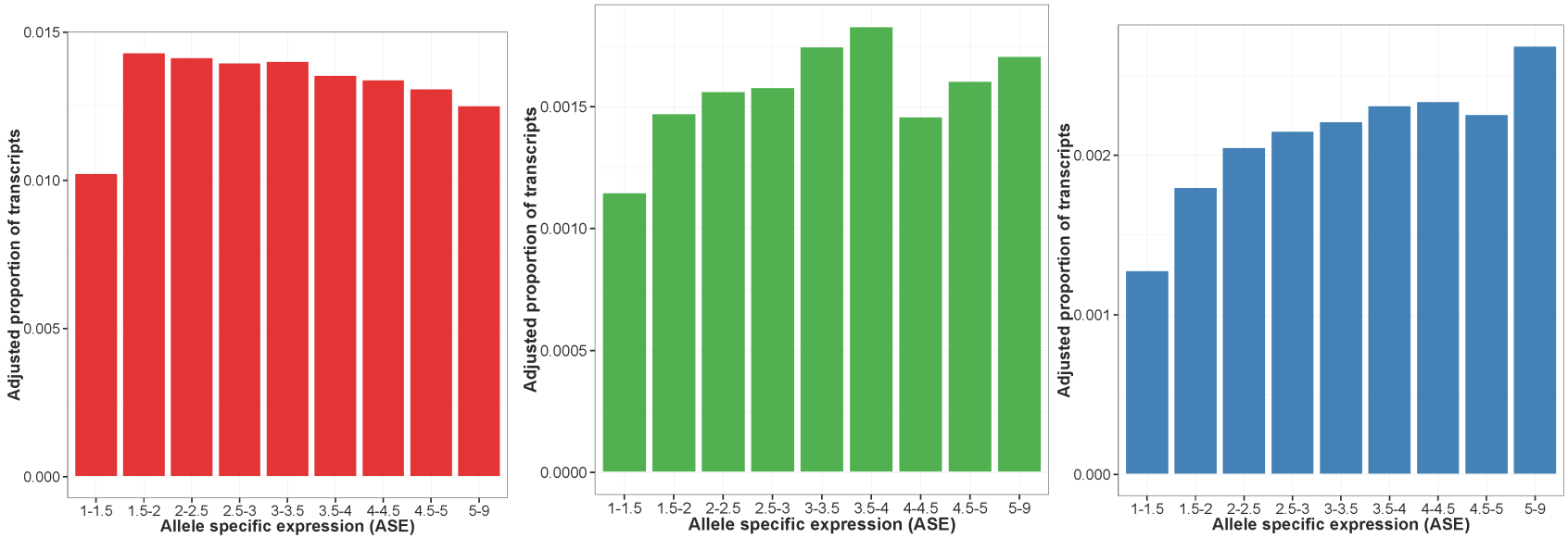


Figure S11. Replication set distribution of allele specific expression (ASE). Distribution of proportion of ASE in transcripts with common (red), rare (blue) or novel (green) noncoding variants in vicinity (+/-20kb from gene) adjusted for average number of SNPs used to calculate ASE.

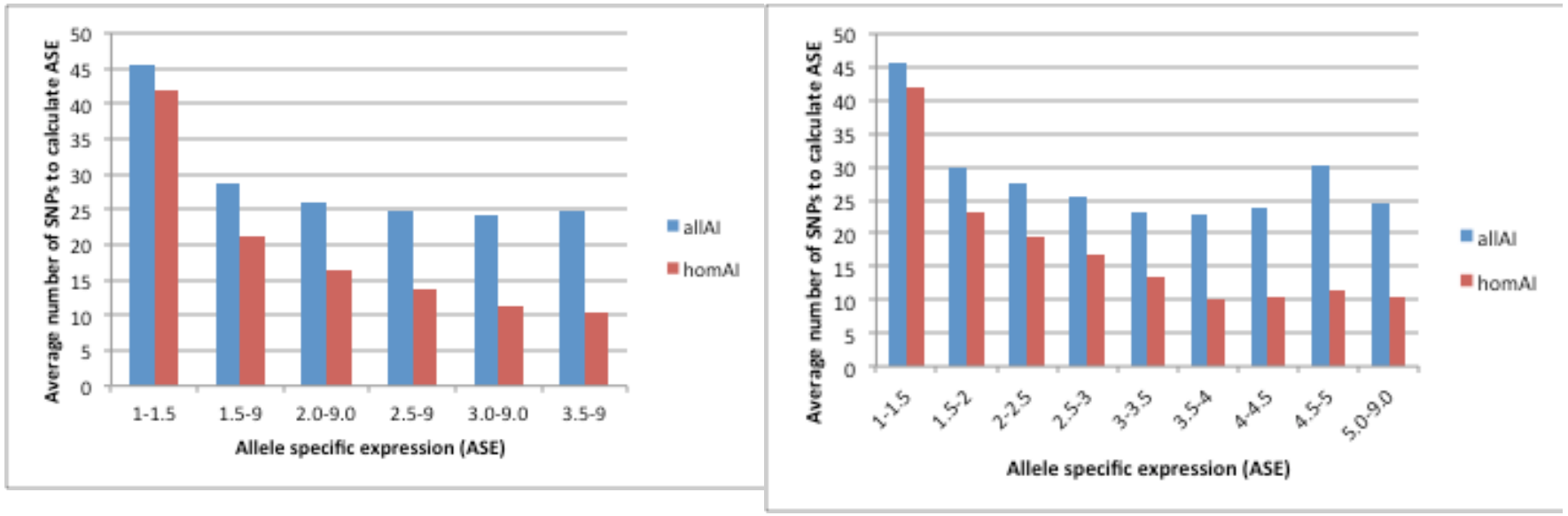


Figure S12. Average number of SNPs used to calculate allele specific expression (ASE) in the replication set. Comparing all transcripts for which ASE was measured (allAI) and transcripts for which the top associated SNP that drives the association across samples is homozygous (homAI).

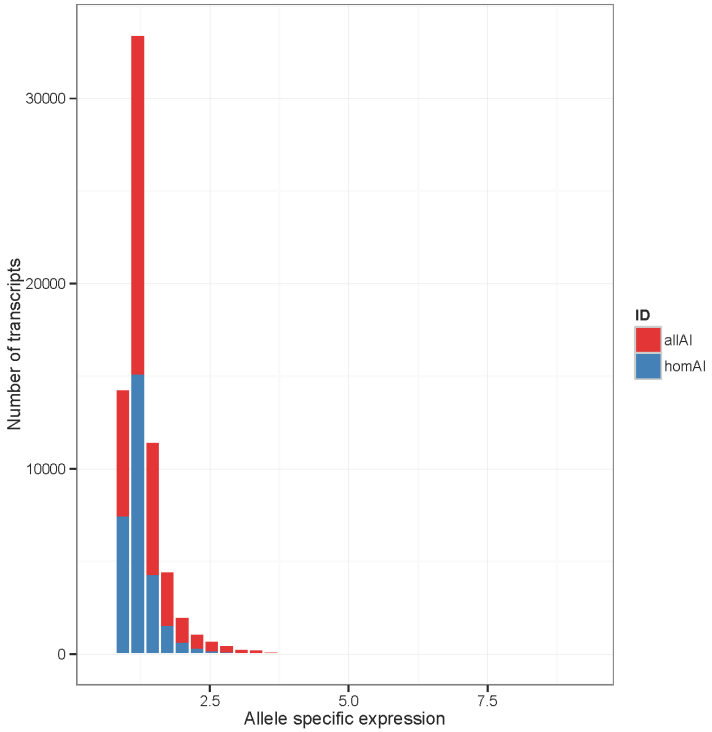


Figure S13. Distribution of allele specific expression of all transcripts and transcripts that did not carry the common allele in a heterozygous state. Histogram of number of transcripts from each category (overlay of the two). Comparing all transcripts for which ASE was measured (allAI) and transcripts for which the top associated SNP that drives the association across samples is homozygous (homAI).

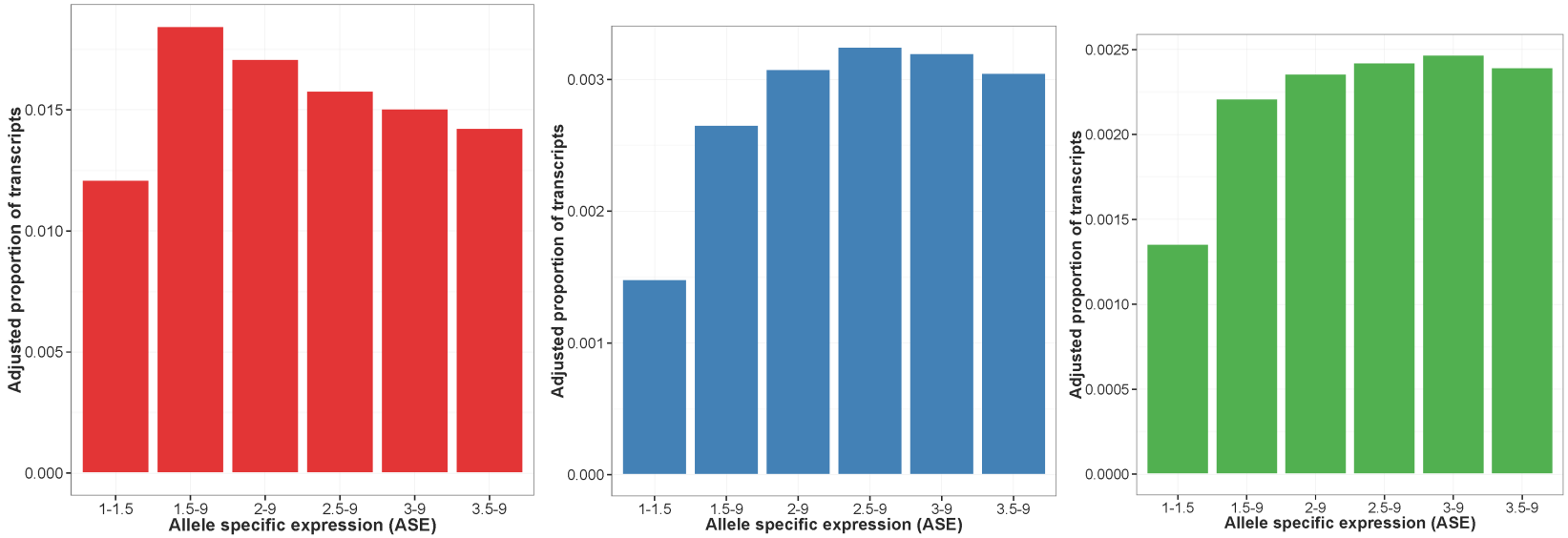


Figure S14. Replication set distribution of Allelic Imbalance (AI). Adjusted proportion of transcripts with common (red), rare (blue) or novel (green) noncoding variants in vicinity (+/-20kb from gene) based on different AI: 1.5 to 9, 2 to 9, 2.5 to 9, 3 to 9 and 3.5 to 9-fold difference. Including only transcripts for which the top associated SNP is homozygous in the sample (homAI).

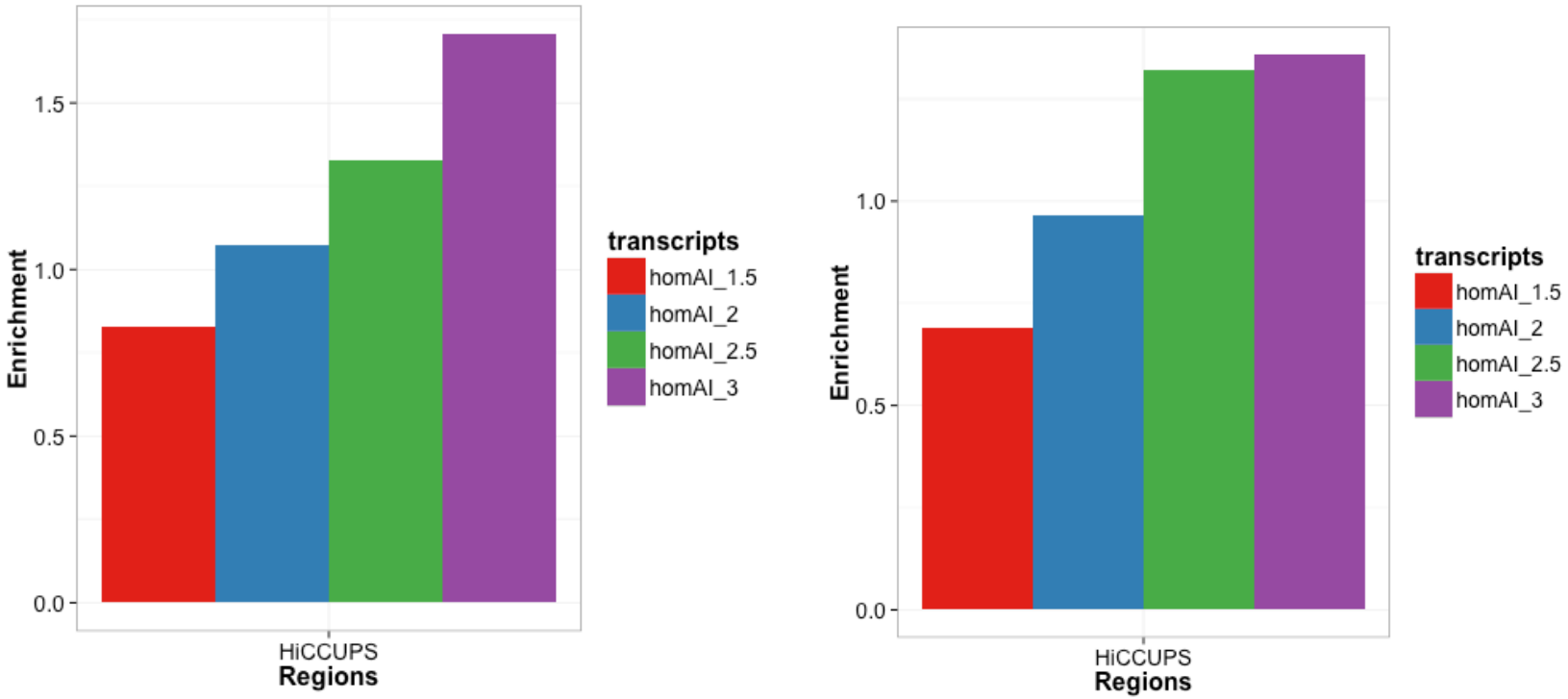


Figure S15. Enrichment between proportion of AI transcripts with rare or novel variants in vicinity compared to AI transcripts with common variants in vicinity in the discovery and replication set. Only including transcripts for which the top associated SNP is homozygous (homAI). We looked at promoter regions as well as regions linked to it by Hi-C. We compare different levels of allelically imbalanced transcripts from 1.5-fold to 3. Results from both the discovery set (left) and replication set (right) are shown.

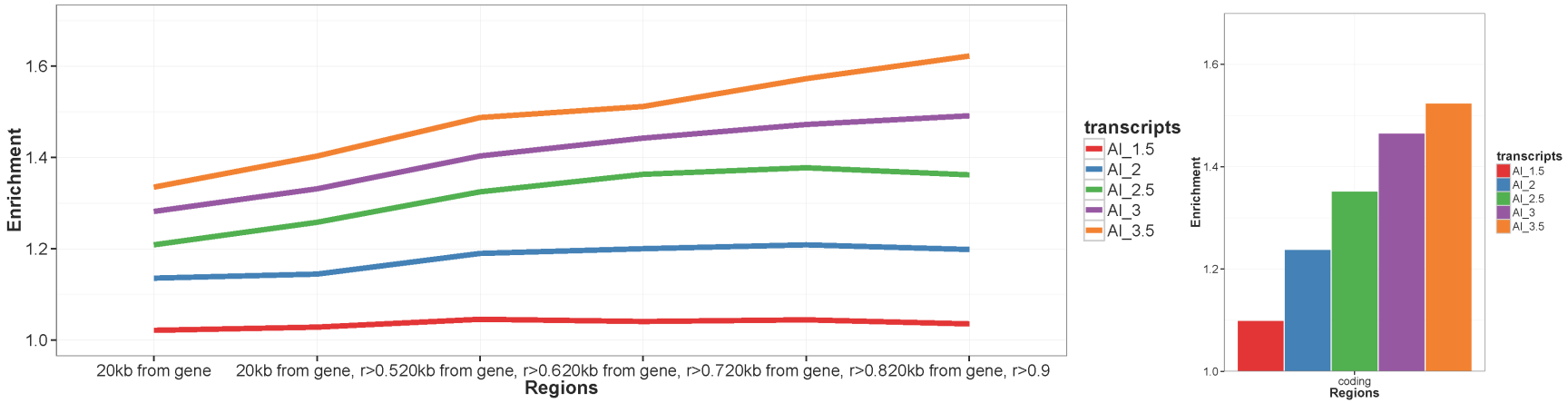


Figure S16. Enrichment between proportion of AI transcripts with rare or novel variants in vicinity compared to AI transcripts with common variants in vicinity in the replication set. Including all transcripts (allAI). We looked at coding (histogram) vs noncoding variants around the genes (+/-20kb from gene) and in DHS regions correlated with the promoters (Pearson correlation r>0.5 to 0.9). We compare different levels of allelically imbalanced transcripts from 1.5-fold to 3.5.

Chapter 3

Preface: Bridging Text between Chapters 2 and 3

Chapter 2 described the design and potential efficiency of the Immune-genetics sequencing that we developed to study rare and low-frequency coding and non-coding regulatory variants in autoimmune and inflammatory complex traits. In this chapter, we used it on the Saguenay–Lac-Saint-Jean asthma familial cohort, which is a founder population. Using this population allowed us to explore two questions: 1) is the SLSJ population enriched in deleterious variants as seen in other founder populations and 2) what are the impact of rare and low-frequency variants on asthma and allergy-related traits and can we identify new genes associated to the traits.

Chapter 3: Exploring rare and low-frequency variants in the Saguenay–Lac-Saint-Jean population identified genes associated with asthma and allergy related traits

Running Title: Rare and low-frequency variants in asthma and allergy

Andréanne Morin[1,2], Tony Kwan[1,2], Anne-Marie Madore[3], Maria Ban[4], Jukka Partanen[5], Lars Rönnblom[6], Ann-Christine Syvänen[7], Stephen Sawcer[4], Hendrik Stunnenberg[8], Mark Lathrop[1,2], Tomi Pastinen[1,2,9,*], Catherine Laprise[3,10*]

1. Department of Human Genetics, McGill University, Montréal, Quebec, Canada.
2. McGill University and Genome Québec Innovation Centre, Montréal, Quebec, Canada.
3. Département des sciences fondamentales, Université du Québec à Chicoutimi, Saguenay, Quebec, Canada.
4. Department of Clinical Neurosciences, University of Cambridge, Cambridge, UK.
5. Research & Development, Finnish Red Cross Blood Service, Helsinki, Finland.
6. Department of Medical Sciences, Section of Rheumatology, Uppsala University, Uppsala, Sweden.
7. Department of Medical Sciences, Molecular Medicine and Science for Life Laboratory, Uppsala University, Uppsala, Sweden.
8. Department of Molecular Biology, Faculty of Science, Radboud University, Nijmegen, the Netherlands.
9. Center for Pediatric Genomic Medicine, Kansas City, Missouri, USA
10. Centre intégré universitaire de santé et de services sociaux du Saguenay, Saguenay, Quebec, Canada

*Corresponding authors

All authors declare no conflicts of interest

This work was supported by Laprise and Pastinen CIHR grant.
AM was supported by the Fonds de Recherche du Québec – Santé (FRQS) doctoral training award.
JP was supported by grant #288393 from the Academy of Finland, by Finnish Funding Agency for Innovation (TEKES) to Salwe GetItDone program and the Finnish government VTR funding.
Collection and sequencing of the Swedish populations samples was supported by a grant from the Knut and Alice Wallenberg Foundation (to A-C S and LR).

3.1 Abstract

The Saguenay–Lac-Saint-Jean (SLSJ) region is located in northeastern Quebec and is known for its unique demographic history and founder effect. Since founder populations are enriched in private variants, we first assessed differences in variant distribution and characteristics between this population and the Finnish founder population along with three other European populations (Sweden, United Kingdom and France). We then explored the advantages of using the SLSJ population in the study of rare coding and noncoding regulatory variants in complex traits such as asthma and allergies. We used targeted sequencing of coding and non-coding regulatory regions of immune cells on 149 trios from the SLSJ asthma familial cohort and on samples from the four European populations. Although the founder populations do not appear to have more rare or deleterious variants, we observed that a larger proportion of private variants reached higher frequencies and that low-frequency variants appear to be more deleterious. Thus, a substantial number of variants private to these populations can be tested in the context of complex traits. Genotypes were then inferred and imputed for the rest of the SLSJ cohort (1 214 samples) and used in single variant association and gene-based tests on asthma and allergy related-traits: eosinophil percentage, immunoglobulin (Ig)E levels and lung function. Using a founder population like the SLSJ allowed us to identify four new genes associated with asthma and allergy-related traits. This may help better understand the genes and pathways implicated in the development of the pathophysiology.

Keywords: asthma, allergy, rare and low-frequency variants, founder population

3.2 Introduction

The Saguenay–Lac-Saint-Jean (SLSJ) region is located in northeastern Quebec and is known for its unique demographic history and founder effect, characterized by several population bottlenecks followed by rapid expansion [186]. Founder populations have been useful to successfully identify rare variants associated with different complex traits [68, 183]. The advantage resides in their homogeneity and the genetic drift resulting in distinctive allele frequencies. In fact, deleterious alleles could overcome their selective disadvantage by reaching higher frequencies in the population. Other studies also suggested that an advantage of founder populations could be their enrichment in deleterious variants [68, 77, 228], which was shown in French Canadians [78] and Finnish populations [79]. However, several studies have challenged these observations [80-82].

Studying rare and low-frequency variants in complex traits is a step forward from the investigations of common genetic variants and provided additional information on the underlying biological mechanism. Despite the important contribution of common variants, we now know that they only explain part of the picture about the genetic basis of complex traits [14]. Additional genetic burden may exist in the low-frequency and rare spectrum of genetic variation that have been explored only more recently.

The first goal of this paper is to assess differences in variant distribution and characteristics of the SLSJ founder population by comparing it to the Finnish founder population and three other European populations from Sweden, United Kingdom, France. Since SLSJ is part of the French-Canadian population, we wanted to follow up on previous results and see if we could observe the same enrichment of deleterious variants

in this population [78]. We will then explore the impact of rare and low frequency variants in asthma and allergy-related traits in the SLSJ asthma familial cohort [186]. Asthma and allergy related traits are common diseases with an important genetic component that remains unexplained [121]. Rare and low-frequency variants have been previously studied to better understand the genetic basis of asthma and allergy related traits [27, 67, 158]. Although they do not explain a large part of the missing heritability of the disease they do take part into its architecture [27]. Rare variants exploration identified previously associated genes (ex: *GSDMB [27]*, *IL33 [67]*) and new ones (ex: *GRASP* [27]), highlighting the importance of studying their role in the context of complex traits.

In this study, we take advantage of the well-described population and availabilities of multiple, related continuous phenotypes. Asthma includes multiple subphenotypes and endotypes, therefore we limited our asthma definition and focused on related continuous phenotypes could help us identify variants associated with the disease. We restricted to IgE levels, eosinophils (Eos) percentage and lung function. The aim was to identify new genes/variants associated with the traits to help further understand biological mechanisms and pathways related to asthma or allergic diseases. We were able to identify one low-frequency variant and two genes associated with Eos percentage and serum IgE levels. These results provide rationale for sequence-based association studies in founder populations to identify variants that may be missed using genotyping chip and/or larger heterogeneous populations.

3.3 Material and Methods

Samples

We sequenced 149 trios (447 samples) from the SLSJ asthma familial cohort [186] using a custom targeted capture panel developed by our group [229] followed by next-generation sequencing. This panel covers around 3% of the genome, including coding and non-coding immune regulatory regions [229]. For the first part of this study, we paired the non-related parents to samples from four other populations: Finland (FINN), France (FR), United Kingdom (UK) and Sweden (SWE). 93 samples from the five different populations were included and paired by mean coverage to avoid bias. All studies received ethic approbation from their respective ethic committees. To analyze the impact of rare variants on lung function (Forced vital capacity (FVC), Forced expiratory volume in one second (FEV_1), and Tiffeneau-Pinelli index (FEV_1/FVC)), serum IgE levels and Eos percentage we used well-described samples from the SLSJ asthma familial cohort (see Table 1 for Clinical description and[1] for Recruitment details). This cohort includes 1 214 individuals from 271 families[1]. The 149 sequenced trios are part of larger families, of which 110 siblings had the sequenced inferred and imputation was performed in the rest of the cohort (see *Inference and Imputation* section). Lung function, serum IgE levels and differential white blood cell counts are all described in Laprise 2014.[1] The study was approved by the *Centre intégré universitaire de santé et de services sociaux de Saguenay* ethics committee. All subjects gave informed consent.

Capture and Sequencing

Samples from the 149 trios from the SLSJ asthma familial cohort as well as the 372 samples from the four other European populations were all sequenced using our custom capture panel. See Morin *et al.*[16] for details about panel description, capture and sequencing. To remove any potentially related individuals, we performed an identity-by-descent estimation using the method of moments[17] and a principal component analysis using the SNPRelate R package[18] (Supplementary Figure S1a). We also used heterozygous/homozygous proportion to identify and remove outliers (Supplementary Figure 1b-d). We also removed paired samples and kept 76 samples per population for a total of 380 samples.

Variant calling and filtering

We aligned reads to the Genome Reference Consortium Human genome build 37 (GRCh37) using bwa 0.7.6a. and we called variants using HaplotypeCaller v3.2 (GATK). We performed merge calling for all selected samples for part one together (465 samples) and the 149 trios independently. For the 149 trios, a mean coverage of 37.6x was obtained. We compared sequencing to genotyping data (see *Genotyping* section) as a quality control using heterozygous and biallelic variants that were both in capture region and on the genotyping chip. We remove seven samples that had a concordance of less than 95% (440 samples remained). We also used the comparison between sequencing and genotyping to set our "true variant cut-off"; at read depth (dp) >10x, genotyping quality (gq) >35, we observed and accuracy of >95% and a sensitivity of >95%. We assessed Mendelian errors using VCFtools [232] and a parent/proband pair was excluded due to high error rate. Mendelian errors were replaced by missing values. Variants included in

both sample sets met these criteria: 1) Fall within the targeted regions, 2) biallelic sites, 3) dp >10x and gq >35 in at least 90% of the samples. We also remove the HLA region, as it will be analyzed independently. A total of 192,228 variants (178,613 SNVs and 13,614 indels) were included in the first part of this study and summarized in Table 2. All five populations had a mean coverage of 28X (18-52X). For the trios, the sequenced was inferred and imputed in the rest of the cohort (see *Inference and Imputation* sections). Functional annotation was performed using SNPeff [220] and selectively constrained variants were identified with the Genomic Evolutionary Rate Profiling (GERP++ score)[218].

Genotyping

Samples from the SLSJ asthma familial cohort genotyping details about chip, DNA extraction samples and variant filtering are described in Laprise 2014 [186] and Moffat *et al.* 2010 [121]. Genotyping data was used for quality cut-offs assessments, inference and imputation of the sequence in the whole cohort.

Inference and Imputation

Genotype phasing was performed using SHAPEIT v2 and duoHMM [233-235] to consider familial structures. Pre-phasing was done using the trios (440 samples) using merged sequencing and genotyping data as well as on the whole cohort using only genotyping data (1 214 samples). We inferred the sequence in the non-sequenced siblings that were part of the same families as the trios. Chromosomes were separated by breakpoints that were identified using duoHMM and NUCFAMTOOLS [236]. We were

able the reassemble the sequence in the siblings using informative markers (heterozygous in one parent and homozygous in the other) and we inferred on average 98.72% (96.89-99.33%) of the sequence at an accuracy of 99.67% (99.39-99.86%). We used the parental haplotypes (294 samples) as the reference panel to reduce the number of duplicated haplotypes and we imputed the sequence using IMPUTE2 [237] in the whole cohort. Missing genotypes from the inference were added using the imputed data. A total of 112 154 variants were imputed with a mean accuracy of 99.7% (98.55-99.98%). The imputation accuracy was measured by comparing imputed probands to their sequenced data. Missing genotypes from the inference were added using the imputed data. We then retained only variants that had an imputation quality of >0.8 for a total of 112 083.

Association testing and gene-based burden analysis

We explored variants associated with the five different phenotypes using EPACTS software (Efficient and Parallelizable Association Container Toolbox: http://genome.sph.umich.edu/wiki/EPACTS). We performed a mixed model association called EMMAX (Efficient Mixed Model Association eXpedited) that accounts for sample structure (population structure and relatedness (kinship coefficient)). EMMAX supports single variant association tests and different burden tests (Combined Multivariate and Collapsing (CMC) [238] and Sequence kernel association test (SKAT) [239] methods). Sex and age were used as covariates as well as height for lung function assessment. We performed single variant analyses on the low-frequency variants (MAF >=0.01 and <0.05) that were not common in UK10K [65] or 1000 genomes (1KG) [55]. For the gene-based test, we collapsed rare and low-frequency variants per genes, including 20KB

around it and only included gene region that had at least two variants for a total of 14 646 ($p<3.4e-6$ or a FDR 5%).

DNA methylation

We performed DNA methylation association study on a subset of individuals from the SLSJ asthma cohort in whole blood (167 samples) and isolated Eos from blood (24 samples). Eos cell isolation was performed as described in Ferland *et al.* [240] and DNA extraction and sodium bisulfite conversion were described in Liang *et al.* [197]. Methylation levels were assessed using the Infinium HumanMethylation450 BeadChip array (Illumina, San Diego, CA, USA). Normalization steps were described in Morin *et al*.[241] and we applied a robust linear regression model including age and sex as covariates as well as cell type composition for whole blood. We performed association for serum levels, FEV_1/FVC and asthma for the genes associated with Eos percentage. To assess if results were obtained by chance, we resample randomly 1000 times the same number of CpG as observed in the vicinity of each gene and observed if we get the same number of CpG that reached $p<0.05$.

3.4 Results

Founder population are enriched with private low-frequency and common variants

For the first goal, we compared SLSJ samples we paired to samples from the four other European populations (76 samples per population, 380 total). Given the small number of samples that were included in the study, we were only able to assess singletons (variants seen once) in the rare variant spectrum. We first observed a smaller number of variants in both founder populations (Table 1), which was reflected in the number of singleton variants (Supplementary Table 1 and Supplementary Figure 2). A smaller proportion was observed in the SLSJ (21.6%) and FINN (19.4%) populations compared to other populations (28%, 24% and 25% for FR, SWE and UK respectively, Figure 1a), which may reflect the bottleneck events that characterize these populations. Similar proportions of low-frequency variants were observed across populations. When looking at the number of variants per samples, we observed a higher number of low-frequency variants in the founder populations (ANOVA $p<2e-16$; Figure 1b). Another interesting aspect resides in the private variants (seen in one population) that reach higher frequency in the SLSJ and Finnish populations (Figure 1c-d). We then wanted to assess if we observed an enrichment of functional variants in the SLSJ and Finnish founder populations. We did not observe any larger proportion of functional (non-synonymous, loss of function (LoF) or GERP++>1) rare, low-frequency or common variants in the two founder populations (Supplementary Figure 3). However, we observed a tendency of higher non-synonymous/synonymous ratio in the low-frequency ($p<0.05$) and singleton portion of the founder populations compared to the others (Supplementary Figure 4a). The tendency was also observed when looking at the per sample distribution ($p<5e-10$, Supplementary Figure 4c-d). We also looked at the enrichment of LoF, non-synonymous and

synonymous variants of the two founder populations compared to the population from FR, SWE and UK (Supplementary Figure 5). We observed an enrichment of non-synonymous low-frequency variants when comparing the SLSJ to FR and UK and when comparing the FINN to FR. The largest enrichment was observed for LoF variants; however, this did not reach significance given limited sample size. Similar pattern was observed when looking only at private variants (Supplementary Figure 8). We also looked at the average GERP++ score per sample distribution and found no difference between the populations (Supplementary Figure 6). However, focusing on low-frequency variants we observed a higher GERP++ score in the founder population compared to the French and UK (Supplementary Figure 7 and 9). These results were also reflected in the non-synonymous variants but not the synonymous or the LoF ones. Overall, we observed a higher proportion of private variants reaching higher frequencies in the founder populations. We also observed a tendency of enrichment for more deleterious variants in the founder population, especially in the low-frequency spectrum.

Assessing the impact of rare and low-frequency variants in asthma and allergy related-traits: Single-variant association analyses

We used the SLSJ asthma familial cohort that comprises 1 214 samples from extended families. We used data from the unrelated parent of the 149 sequenced trios to infer and impute the sequence in the rest of the cohort (see Methods section). We assessed the individual effect of the coding and non-coding low-frequency (MAF between 1% and 5%) on five different asthma and allergy related traits (serum levels, FEV_1, FVC, FEV_1/FVC and Eos percentage). Both Manhattan and qqplot (lambda from 0.98 to 1.075)

can be found in Supplementary Figure 10. We observed a significant association (p<3.26e-6) with Eos percentage for a SNV (rs1386931) located in the 3'UTR of *CXCR6* and in the intron of the *FYCO1* genes (Table 3). We also observed another SNV reaching suggestive significance (p<1e-5) with serum levels and located in the intron of the *NRP2* gene (Table 3). No variants were identified for lung function. We then tested if the two SNVs reaching P<1e-5 were also associated with other the other traits as well as asthma, atopy, allergic asthma, rhinitis and atopic dermatitis (Supplementary table 4). The SNV in the *NRP2* intron associated with serum IgE levels was also marginally associated with atopy and allergic asthma. The significantly associated SNVs for Eos percentage had p<0.05 to serum IgE levels and atopic dermatitis. Both of the SNVs were also found in 1KG and UK10K. The SNV associated with Eos percentage (rs1386931) has a higher minor allele frequency in SLSJ (0.043) compared to the one observed in 1KG (0.021) and UK10K (0.019). The other variant (rs849558) had slightly higher frequency in SLSJ (0.019) compared to UK10K (0.011).

Gene-based analyses

We then used gene-based test that combines variants with MAF<5% in a region to get more power to detect association. We used SKAT and CMC tests as implemented in EPACTS on serum IgE levels, FEV_1, FVC, FEV_1/FVC and Eos percentage. We combined rare and low-frequency variants together by gene including 20kb region around it. Manhattan and qqplots (lambda values range from 1 to 1.05 for CMC and 1.02 to 1.13 for SKAT) for each phenotype and the two tests are presented in Supplementary Figures 11 and 12. Genes with p<3.4e-6 for one of the two tests are listed in Table 4 and the ones

that reached P<1e-5 in Supplementary Table 5. We also reported the lead SNVs that were identified when running the tests again removing one variant at a time. Two genes were significantly associated with Eos percentage (*MRPL44*) and serum IgE levels (*NRP2*). One SNV lead each association: a rare one for *MRPL44* (rs76568361) and a low-frequency one for *NRP2* (rs849558). The latter almost reached significance in the single variant association test (Table 3). The lead variant for *MRPL44* is intergenic and is also located close to *SERPINE2* (SKAT p= 1.47e-5) and has much smaller frequency in the 1KG European populations and in UK10K (Supplementary Table 5). Moreover, we identified four marginally associated genes (p<1e-5, Supplementary Table 5): two with Eos percentage (*SHMT1* and *SMCR8*) and two with FEV_1/FVC (*CCDC126* and *CLK2P*). The lead SNV was the same for *SHMT1* and *SMCR8,* a rare missense variant (rs79875842) located in the latter gene. The same situation was observed for the FEV_1/FVC genes where the two lead SNVs were missense variants (rs73077128 and rs146336907) for the pseudogene *CLK2P*. The variant rs146336907 was not observed in 1KG or in UK10K (Supplementary Table S5).

DNA methylation in associated genes

To support our associations observed from single variants and gene-based association test, we performed DNA methylation analyses of CpG located +/- 20KB from the associated genes (n= 182, Supplementary Table 6) in whole-blood (167 samples) or isolated Eos (24 samples). We observed nine CpGs with p<0.05 for serum IgE levels in isolated Eos in the *NRP2* gene, which is larger than expected (4.8), as well as a CpG

reaching significance located in the gene intron (p=3.7e-4, n=49). This result supports the importance of *NRP2* gene association with serum IgE levels.

3.5 Discussion

In our study, we explored two aspects of rare and low-frequency variants in the SLSJ population: their distribution and the enrichment of deleterious variants as well as their impact on asthma and allergy related traits. For the first goal, we observed a smaller proportion of private variant found in the SLSJ and FINN founder population, which does not supports what was previously observed in the French Canadian population [78], but does for the FINN one [79]. However, our results reflect both studies as we observed a tendency for enrichment of deleterious variants in the two founder populations, especially in the low-frequency spectrum of variants. We did not reach significance in all functional variant categories probably due to limited sample size. Another interesting result was the larger proportion of private variants reaching higher frequencies in the founder populations, reflecting the genetic drift of the founder population. The strength of our assessment resides in comparing five populations, including two founder ones, all processed the same way and paired based on mean coverage.

We were able to identify one significant low-frequency variants associated with Eos percentage and a suggestively significant one associated with serum IgE levels. They were both non-coding highlighting the importance of exploring these regions to understand complex traits. The first one is located in the intron region of the *FYCO1* gene and in the 3'UTR of the *CXCR6* gene. *FYCO1* encodes for a protein that plays a role in the transport of autophagic vesicles [242]. It was never associated with eosinophil

percentage in the past or to any asthma or allergy relate trait. However, autophagy process has been previously linked to asthma [243, 244] and genes having an important role in it were associated with asthma before [191, 245, 246]. *CXCR6* encodes for a chemokine receptor expressed at the surface of multiple immune cells and was previously linked to asthma and Th2 inflammation in the lung [247]. The second suggestively associated variant is located in the intron of the *NRP2*, which is a transmembrane receptor implicated in multiple processes, including in the immune system for antigen presentation, phagocytosis and cell-cell interaction[41]. This gene was also associated in the gene-based test; being led by the same associated SNV and was supported by DNA methylation association. We also identified *MRPL44* gene associated with eosinophil percentage and known to be implicated in protein synthesis in mitochondria. The mitochondria play an important role in Eos apoptosis and survival [42]. Moreover, the lead SNV for *MRPL44* is located in the promoter region of the *SERPINE2* gene, for which the association was suggestive (p=1.47e-5). *SERPINE2* is a serine protease inhibitor and is a known susceptibility gene for chronic obstructive pulmonary disease (COPD) [248], emphysema [249] and asthma [250]. However, no link has been observed with Eos so far. The other four suggestively associated genes using gene-based test, actually pointed to two genes: *SMCR8* associated with Eos percentage and the *CLK2P* pseudogene associated with FEV_1/FVC. *SMCR8*, just like *FYCO1,* appears to potentially regulates the transcription of autophagy related genes [251].

3.6 Conclusion

In this study, we first showed that founder populations appear to be enriched in deleterious low-frequency variants. We then pursued testing the impact of rare and low-frequency variants in asthma and allergy related-traits. Using our custom capture panel on the SLSJ founder population we identified coding and non-coding rare and low-frequency SNVs associated with Eos percentage, serum IgE levels and FEV_1/FVC. One of the lead SNV in the gene-based test was private to the SLSJ population highlighting the importance of using sequencing data in founder population to identify new genes associated with complex traits. Other SNV also presented marginally higher frequency compared to the European population. We note that quantitative rather than discrete variation reached significance underscoring importance of intermediate phenotypes in complex traits. We also demonstrate the importance of addressing the non-coding regions of the genome by using sequencing studies, as three of the variants identified were non-coding. Overall, we showed the advantage of using a well-described founder population and the importance of assessing non-coding regions to better decipher the genetic basis of complex traits.

3.7 Acknowledgements

We thank all families for their participation. This work was supported by Laprise and Pastinen CIHR. AM was supported by the Fonds de Recherche du Québec – Santé (FRQS) doctoral training award. CL is the director of the Asthma Strategic Group of the Respiratory Health Network (RHN), investigator of CHILD Study, and is a member of the AllerGen NCE Inc. CL is the chairholder of the Canada Research Chair in the Environment and Genetics of Respiratory Disorders and Allergies. JP was supported by grant #288393 from the Academy of Finland, by Finnish Funding Agency for Innovation (TEKES) to Salwe GetItDone program and the Finnish government VTR funding. Collection and sequencing of the Swedish populations samples was supported by a grant from the Knut and Alice Wallenberg Foundation (to A-C S and LR).

3.8 Figures and Tables

Table 1. Clinical description of the SLSJ asthma familial cohort

	All samples (n=1214)	All trios (n=447)	Probands[a] (n=149)	Parents (n=298)	Siblings (n=110)
General Characteristics					
M:F ratio	1 :1.17	1:1.1	1:1.3	1:1	1 :1.2
Age, mean (range)[b]	38 (2-96)	36 (3-75)	18 (3-45)	45 (27-75)	14 (2-44)
Age of onset[c]	16 (0-75)	14 (0-64)	7 (0-37)	24 (0-64)	6 (0-44)
Smoking status % (never smoker; former smoker; current smoker)[d]	646;339;209	51;27;22	82;6;12	36;37;27	82;7;11
Clinical descriptive data					
FEV_1, L (SD)[e]	2.93 (0.82)	2.99 (0.76)	2.93 (0.80)	3.01(0.74)	2.93(0.88)
FVC, L (SD)[f]	3.73 (1.02)	3.82 (0.94)	3.71 (1.04)	3.87(0.88)	3.52(1.10)
FEV_1/FVC, % (SD)[g]	94 (9)	72.5 (22.4)	70.8 (29.6)	72.6 (21.2)	77.7 (21.1)
Serum IgE (SD)[h]	471 (1564)	432 (1406)	806 (2309)	251 (501)	276 (404)
Asthma, n (%)[i]	592 (49)	264 (59)	149 (100)	116 (36)	52 (47)
Allergy, n (%)[j]	677 (57)	287 (64)	121 (82)	170 (57)	73 (68)
With asthma, n (%)	433 (36)	206 (46)	121 (82)	90 (30)	37 (35)
Eosinophils [k]					
Count in 1^e9/L (SD)	0.24 (0.22)	0.25 (0.23)	0.32 (0.34)	0.21 (0.15)	0.26 (0.26)
Percentage (SD)	3.6 (2.8)	3.7 (2.8)	4.4 (3.3)	3.3 (2.4)	3.95 (3.4)

[a] Probands are the first family member recruited in the cohort. [b] Mean age calculated for 1 212 subjects, 447 trios members, 149 probands, 298 parents and 110 siblings. [c] Mean age of onset calculated for 567 subjects, 254 trios members, 142 probands, 112 parents and 48 siblings [d]Smoking status was available for 1 194 subjects, 444 trios members, 148 probands, 296 parents and 107 siblings. Ex-smokers are defined as subject who stopped smoking since over one year. [e]The mean forced expiratory volume in 1 s (FEV_1) is measured in L in 925 subjects, 429 trios members, 141 probands, 287 parents and 94 siblings. [f] The mean forced vital capacity (FVC) is measured in L in 908 subjects, 414 trios members, 134 probands, 279 parents and 92 siblings. [g] The mean FEV_1 (L)/ FVC (L) ratio is calculated in % for 907 subjects, 414 trios members, 134 probands, 279 parents and 93 siblings. [h] The geometric mean of immunoglobulin (Ig) E serum concentration is calculated for 996 subjects, 408 trios members, 142 probands, 292 parents and 99 siblings. [i] Present or past documented clinical history of asthma. Asthme phenotype is available for 1207 subjects, 447 trios members, 149 probands, 298 parents and 110 siblings.[j] Allergy is defined as one positive skin prick testing (wheal diameter ≥3mm at 10 min). The allergy phenotype is available for 1193 subjects, 445 trios members, 147 probands, 296 parents and 106

siblings. [k] Cell type profiles are available for 967 subjects, 418 trios members, 137 probands, 283 parents and 98 siblings.

Table 2. Overall description of variants included in the analyses

	All	SLSJ	FINN	FR	SWE	UK
Mean coverage	28.83	28.88	28.86	28.79	28.81	28.82
Ts/Tv	2.24	2.25	2.25	2.24	2.24	2.25
Total SNVs	178 614	103 889	100 696	112 047	106 375	108 511
Total Indels	13 614	7 780	7 789	8 426	8 093	8 426

SLSJ: Saguenay–Lac-Saint-Jean, FINN: Finland, FR: France, SWE: Sweden, UK: United Kingdom, Ts/Tv: transition to transversion ratio, Indels: insertions and deletions.

Table 3. Results of single low-frequency SNV association study with asthma related trait (P<1e-5)

Trait	rsID	Gene	Alleles	MAF	P-value	Effect (SE)
Serum IgE levels	2:206562250 ; rs849558	*NRP2*; intron	T/C	0.019	4.79e-6	-1.243 (0.270)
Eosinophils percentage	3:45989502; rs1386931	*CXCR6*; 3'UTR and *FYCO1*; intron	C/T	0.043	**1.77e-6**	1.534 (0.319)

CXCR6: C-X-C motif chemokine receptor 6, IgE: immunoglobulin E, MAF: minor allele frequency, NRP2: neuropilin 2, SE: standard error, UTR: untranslated region.

Table 4. Genes significantly associated with asthma and allergy related traits

Trait	Gene	n SNPs	n passing[1]	Fraction with rare	P-value SKAT/CMC	Lead SNVs[2]
Eosinophils percentage	*MRPL44*	8	4	0.026	**2.97e-6**/5.74e-5	2:224835223 (rs76568361)
Serum IgE Levels	*NRP2*	4	2	0.001	**3.16e-6**/0.8237	2:206562250 (rs849558)

[1] Number of variants passing threshold (MAF<0.05). [2]Test were ran again removing one variant at a time, lead SNV correspond to the one for which the entire association rely on. CMC: combined multivariate and collapsing test, IgE: immunoglobulin E, MRPL44: mitochondrial ribosomal protein L44, NRP2: neuropilin 2, SKAT: sequence kernel association test, SNV: single nucleotide variations.

Figure 1

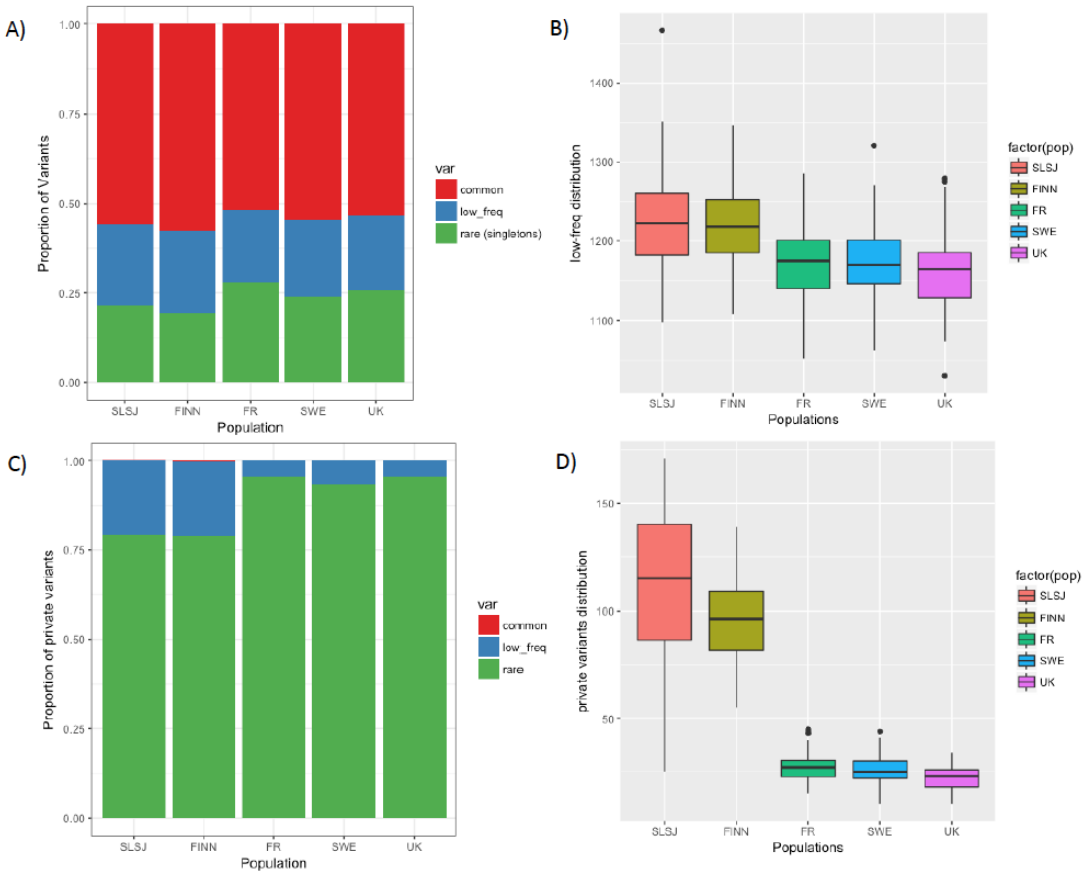


Figure 1. Distribution of variants across founder populations compared to three other European populations. A) Proportion of common (MAF>0.05, red), low-frequency (0.05<MAF>0.01, blue) and rare (MAF<0.01, green) variants in each population. B) Low-frequency variants distribution per sample C) Proportion of private variants of each population, common (MAF>0.05, red), low-frequency (0.05<MAF>0.01, blue) and rare (MAF<0.01, green), D) Distribution of private variants in each population (not including singletons). SLSJ: Saguenay–Lac-Saint-Jean, FINN: Finland, FR: France, SWE: Sweden, UK: United Kingdom

3.9 Supplementary information

Supplementary Table 1. Summary of variants in the five populations

		All	**SLSJ**	**FINN**	**FR**	**SWE**	**UK**
Number of SNVs	All	178 614	103 889	100 696	112 047	106 375	108 511
	Common	58 568	58 223	58 421	58 404	58 285	58 413
	Low-freq	21 949	23 362	22 899	22 362	22 527	22 229
	Rare	98 096	NA	NA	NA	NA	NA
	Singletons	64 834	22 304	19 376	31 281	25 563	27 869
Number of indels	All	13 614	7 780	7 789	8 426	8 093	8 259
	Common	3 981	3 936	3 970	3 998	4 016	4 014
	Low-freq	2 019	1 966	2 087	1 976	2 038	1 947
	Rare	7 614	NA	NA	NA	NA	NA
	Singletons	4 852	1 878	1 704	2 480	2 039	2 298

FINN: Finland, FR: France, indels: insertions and deletions, SLSJ: Saguenay–Lac-Saint-Jean, SNVs: single nucleotide variations, SWE: Sweden, UK: United Kingdom.

Supplementary Table 2. Summary of functional variants in the five populations

		SLSJ	FINN	FR	SWE	UK
All	Synonymous	9 323	9 088	10 197	9 538	9 750
	Non-Synonymous	12 094	11 707	13 249	12 511	12 673
	LoF	459	443	486	456	470
Common	Synonymous	5 217	5 275	5 242	5 211	5 238
	Non-Synonymous	5 508	5 638	5 646	5 576	5 609
	LoF	160	164	166	160	160
Low-frequency	Synonymous	2 120	2 072	2 084	2 107	2 054
	Non-Synonymous	3 089	2 993	2 758	2 904	2 762
	LoF	129	119	95	116	105
Singletons	Synonymous	942	770	1 669	988	1 277
	Non-Synonymous	1 937	1 578	2 990	2 083	2 374
	LoF	110	100	148	104	136

FINN: Finland, FR: France, LoF: Loss of function, SLSJ: Saguenay–Lac-Saint-Jean, SWE: Sweden, UK: United Kingdom.

Supplementary Table 3. Summary of private variants in the five populations

		SLSJ	FINN	FR	SWE	UK
	All	14 909	12 037	20 513	13 839	16 677
	Common	25	51	0	0	0
	Low-freq	3 052	2 493	960	931	779
	Singletons	11 832	9 493	19 553	12 908	15 898
All	Synonymous	1 189	962	1 756	1 057	1 329
	Non-Synonymous	2 409	1 996	3 101	2 223	2 488
	LoF	135	122	152	111	140
Common	Synonymous	0	3	0	0	0
	Non-Synonymous	2	8	0	0	0
	LoF	0	1	0	0	0
Low-frequency	Synonymous	247	189	87	69	52
	Non-Synonymous	470	410	111	140	114
	LoF	25	21	4	7	4
Singletons	Synonymous	942	770	1 669	988	1 277
	Non-Synonymous	1 937	1 578	2 990	2 083	2 374
	LoF	110	100	148	104	136

FINN: Finland, FR: France, LoF: Loss of function, SLSJ: Saguenay–Lac-Saint-Jean, SWE: Sweden, UK: United Kingdom.

Supplementary Table 4. Low-frequency variants reaching p<1e-5 in single variant association and their significance level (p-value) in other traits.

rsID	Gene	Asthma	Atopy	Allergic asthma	Rhinitis	Atopic dermatitis	Serum IgE Levels	Eosinophils percentage	FEV_1	FVC	FEV_1/ FVC
2:206562250; rs849558	*NRP2*; intron	0.7986	**0.0040**	**0.0134**	0.0946	0.0773	**4.79e-6**	0.093	0.1959	0.2538	0.1996
3:45989502; rs1386931	*CXCR6*; 3'UTR and *FYCO1*; intron	0.1658	0.2422	0.7746	0.8691	**0.0353**	**0.0052**	**1.77e-6**	0.3213	0.5572	0.5001

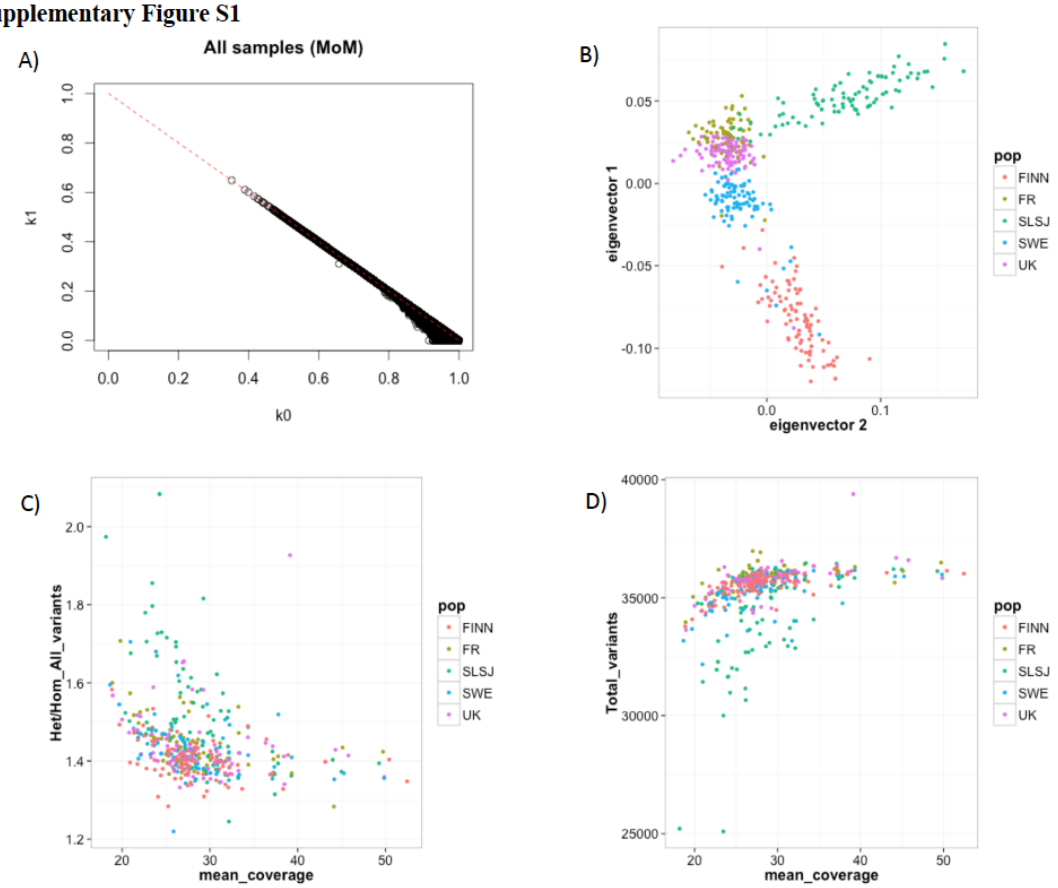
CXCR6: C-X-C motif chemokine receptor 6, FEV_1: forced expiratory volume in one second, FVC: forced vital capacity, FYCO1: FYVE and coiled-coil domain 1, IgE: Immunoglobulin E, NRP2: neuropilin 2.

Supplementary Table 5. Genes reaching p<1e-5 using CMC or SKAT in one of the five asthma and allergy related phenotype.

Trait	Gene	n SNPs	n passing[1]	Frac with rare	P-value SKAT/CMC	Lead SNVs[2]	P-value after removing lead SNV	MAF lead SNV (1KG and UK10K)	P-value single var
Eosinophils percentage	*MRPL44*	8	4	0.026	**2.97e-6**/5.74e-5	2:224835223 rs76568361	0.4425/0.4168	0.0067 (NA and 0.00026)	NA
	SHMT1	7	5	0.017	6.21e-6/3.13e-4	17:18220268 rs79875842	0.7761/ 0.7371	0.0046 (0.0145 and 0.0148)	NA
	SMCR8	8	6	0.018	6.85e-6/3.66e-4	17:18220268 rs79875842	0.8431/ 0.6837	0.0046 (0.0145 and 0.0148)	NA
FEV_1/FVC	*CCDC126/ CLK2P*	3	3	0.022	3.19e-5/4.62e-6	Both SNV: 7:23624887 rs73077128, 7:23625481 rs146336907	8.93e-4 and 0.0018/ 8.65e-4 and 0.0018	0.0036 (0.0106 and 0.0082), 0.0062 (NA)	NA
Serum IgE Levels	*NRP2*	11	8	0.224	**3.16e-6**/0.8237	2:206562250 rs849558	0.0371 (0.021 and 0.019)	0.0191	4.80e-6

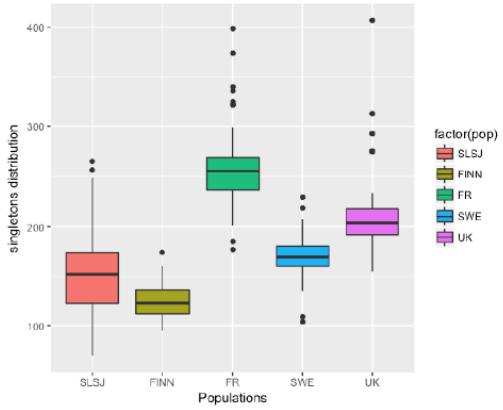
[1] Number of variants passing threshold (MAF<0.05). [2]Test were ran again removing one variant at a time, lead SNV correspond to the one for which the entire association rely on. 1KG: 1000 genomes project, CCDC126: coiled-coil domain containing 126, CLK2P: CDC like kinase 2 pseudogene, CMC: combined multivariate and collapsing test, FEV_1/FVC: Tiffeneau-Pinelli index, IgE: immunoglobulin E, MAF: minor allele frequency, MRPL44: mitochondrial ribosomal protein L44, NRP2: neuropilin 2, SHMT1: serine hydroxymethyltransferase 1, SKAT: sequence kernel association test, SMCR8: Smith-Magenis syndrome chromosome region candidate 8, SNV: single nucleotide variations.

Supplementary Figure S1



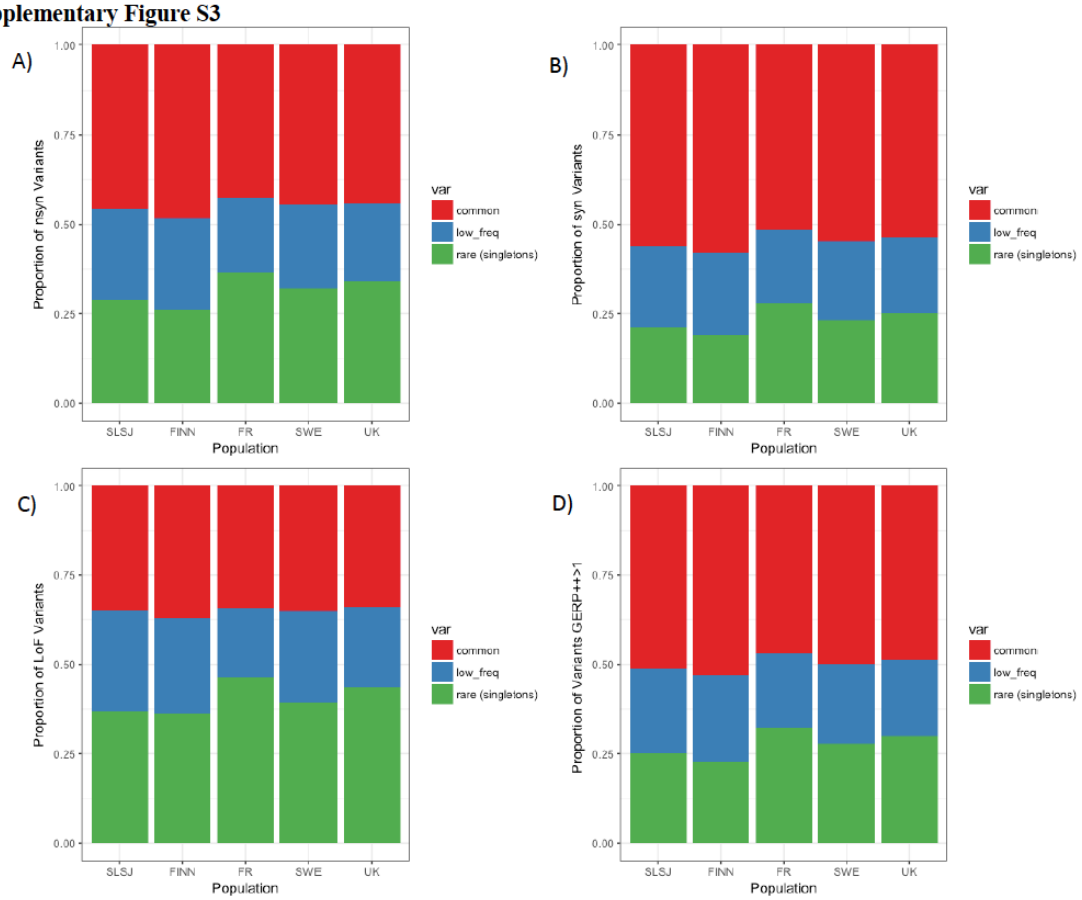
Supplementary Figure S1. Samples selection from the five populations A) Identity-by-descent estimation using method of moments including all 380 samples, B) Principal component analysis (PCA), C) Heterozygous to homozygous proportion compared to mean coverage, D) Total number of variants compared to mean coverage. FINN: Finland, FR: France, SLSJ: Saguenay–Lac-Saint-Jean, SWE: Sweden, UK: United Kingdom

Supplementary Figure S2



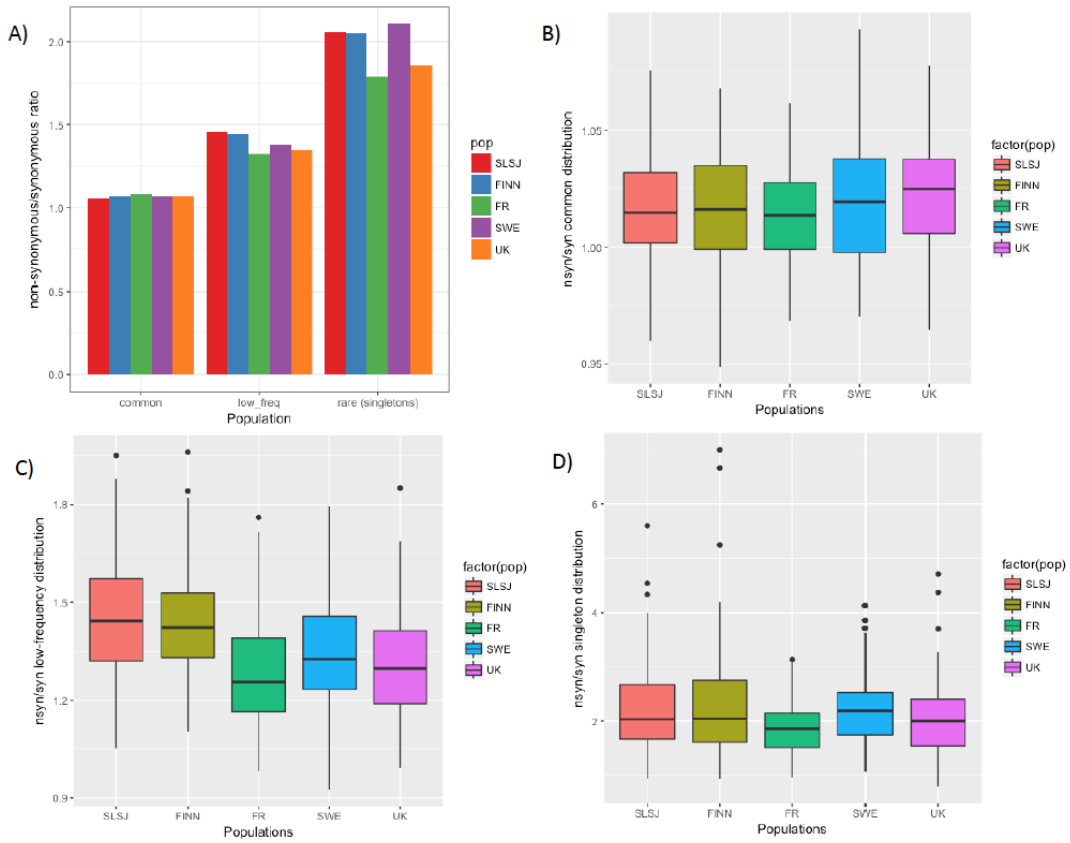
Supplementary Figure S2. Per sample distribution of singletons. FINN: Finland, FR: France, SLSJ: Saguenay–Lac-Saint-Jean, SWE: Sweden, UK: United Kingdom

Supplementary Figure S3



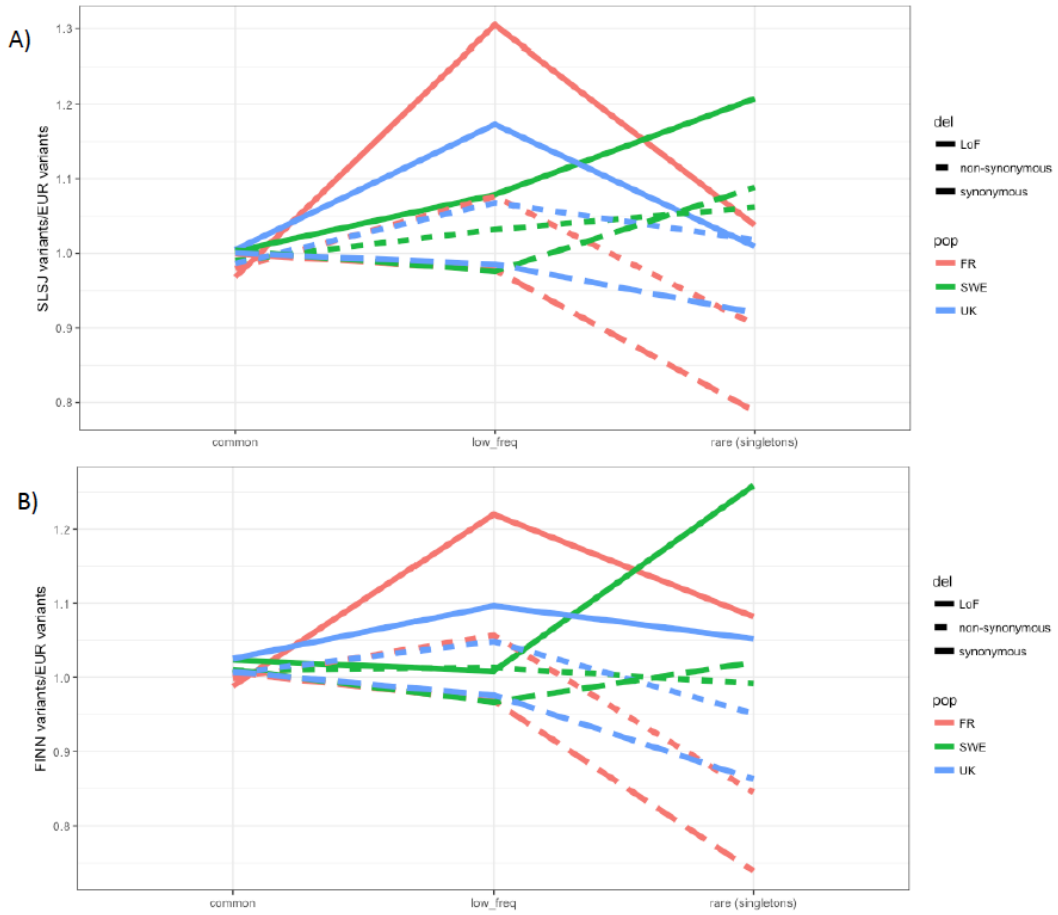
Supplementary Figure S3. Proportion of common (MAF>0.05, red), low-frequency (0.05<MAF>0.01, blue) and rare (MAF<0.01, green) variants in each population. A) non-synonymous, B) synonymous, C) Loss of function and D) GERP++>1 variants from each population. FINN: Finland, FR: France, SLSJ: Saguenay–Lac-Saint-Jean, SWE: Sweden, UK: United Kingdom

Supplementary Figure S4



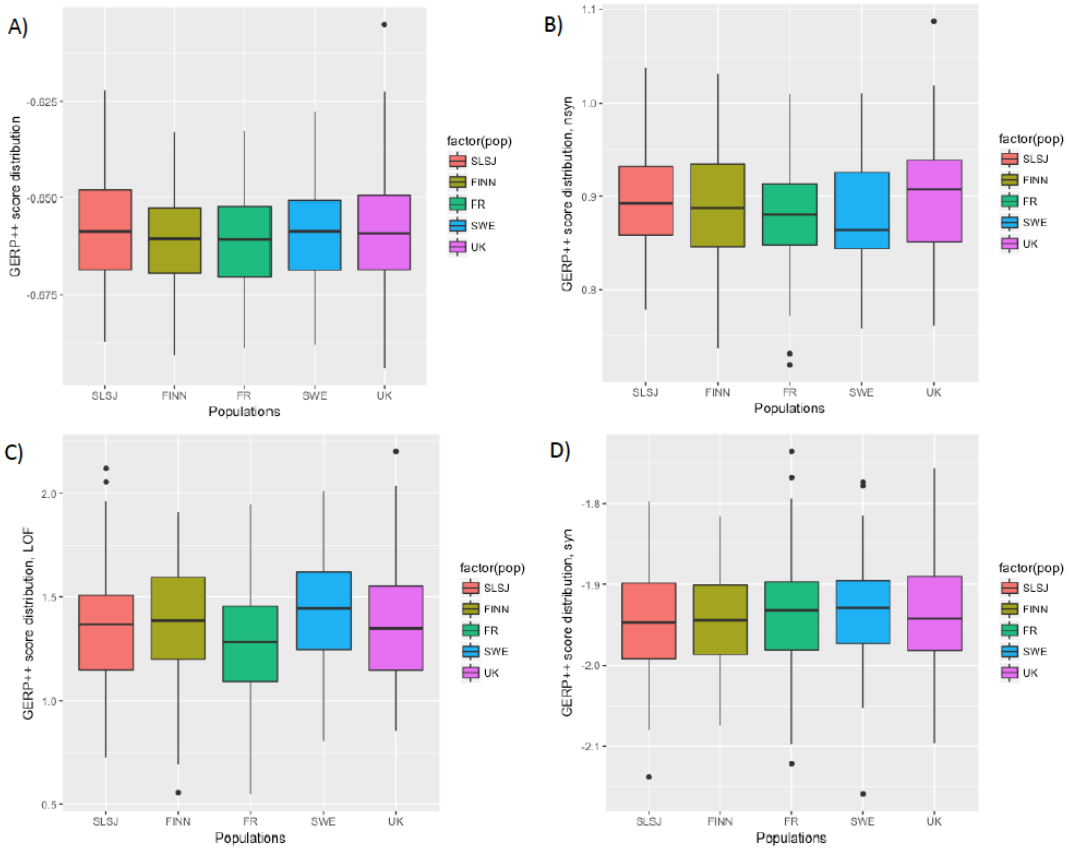
Supplementary Figure S4. Non-synonymous to synonymous ratio A) Total number per population, B) Per sample distribution, C) Low-frequency per sample distribution and D) Singletons per sample distribution. FINN: Finland, FR: France, SLSJ: Saguenay–Lac-Saint-Jean, SWE: Sweden, UK: United Kingdom.

Supplementary Figure S5



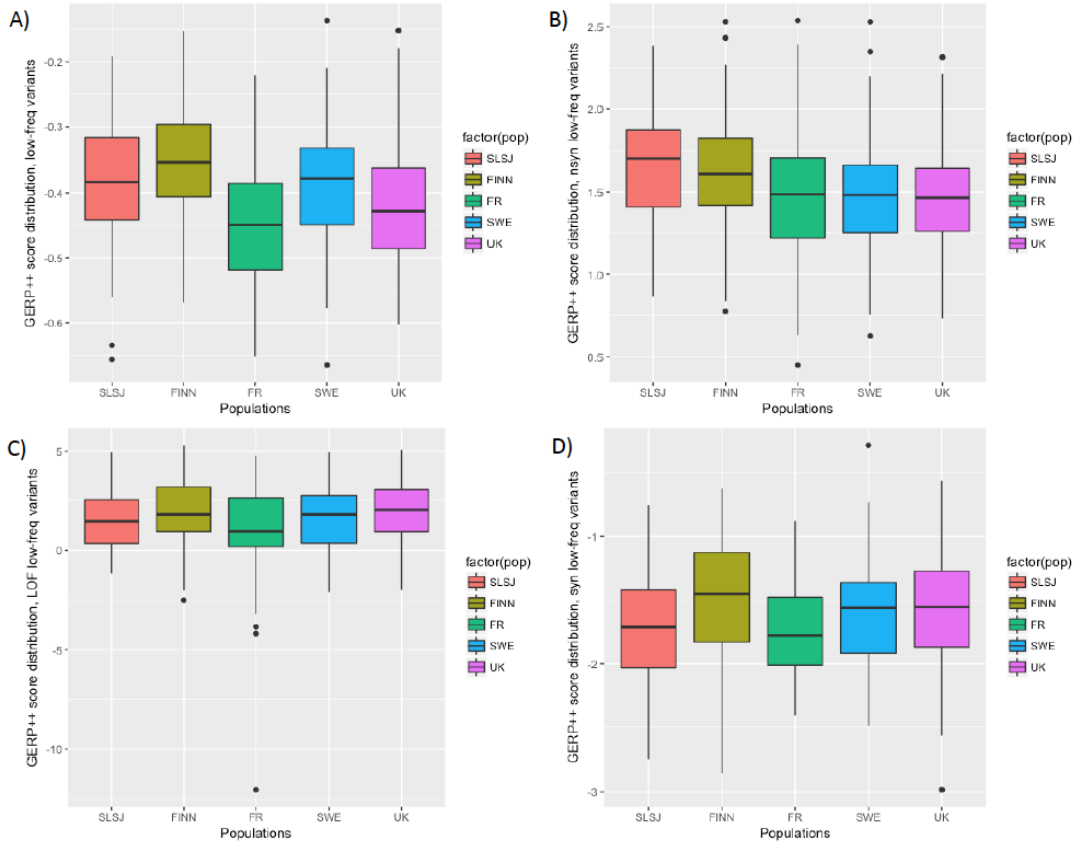
Supplementary Figure S5. Common, low-frequency and singleton variants enrichment in A) Saguenay–Lac-Saint-Jean (SLSJ) and B) Finland (FINN) compared to France (FR), Sweden (SWE) and United Kingdom (UK).

Supplementary Figure S6



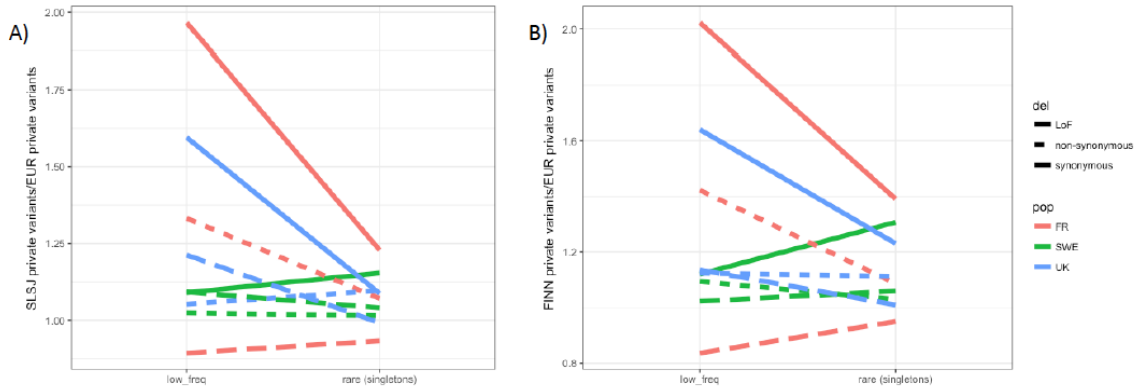
Supplementary Figure S6. Average GERP++per sample distribution per population, A) All, B) non-synonymous, C) loss of Function (LoF) and D) Synonymous. FINN: Finland, FR: France, SLSJ: Saguenay–Lac-Saint-Jean, SWE: Sweden, UK: United Kingdom.

Supplementary Figure S7



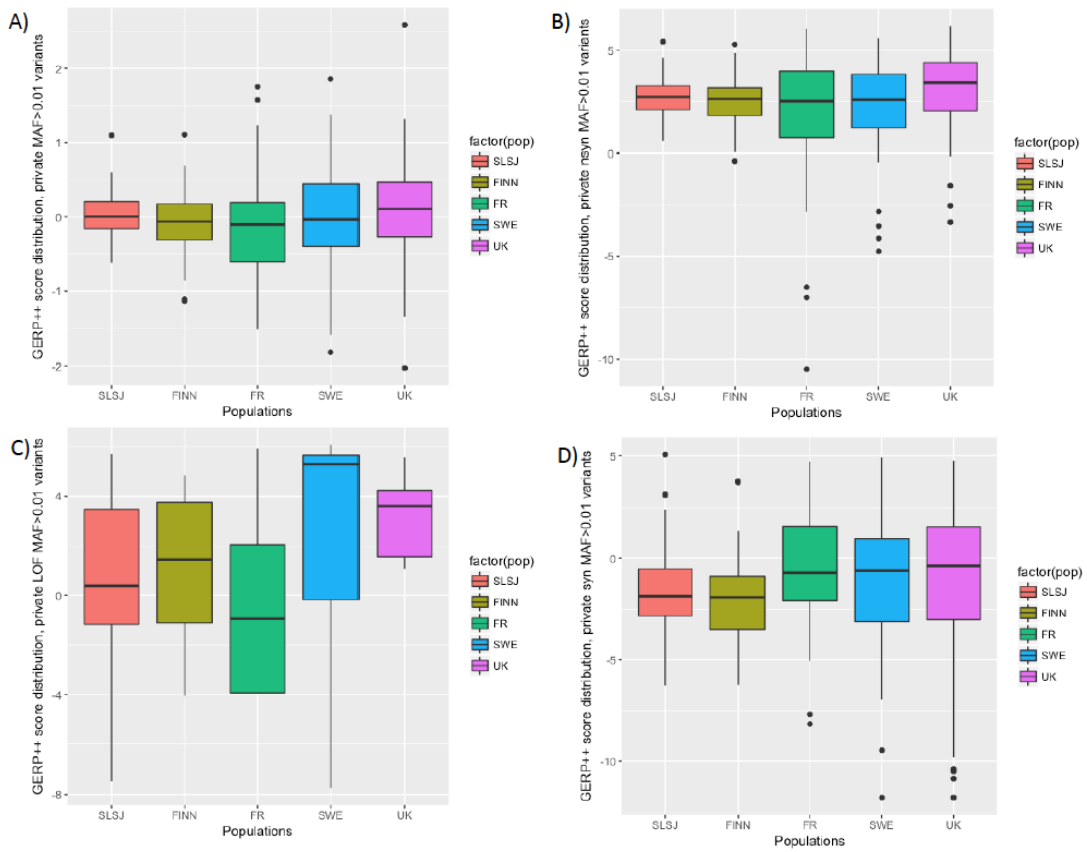
Supplementary Figure S7. Average GERP++per sample of low-frequency variants per population, A) All, B) non-synonymous, C) loss of Function and D) Synonymous. FINN: Finland, FR: France, SLSJ: Saguenay–Lac-Saint-Jean, SWE: Sweden, UK: United Kingdom.

Supplementary Figure S8



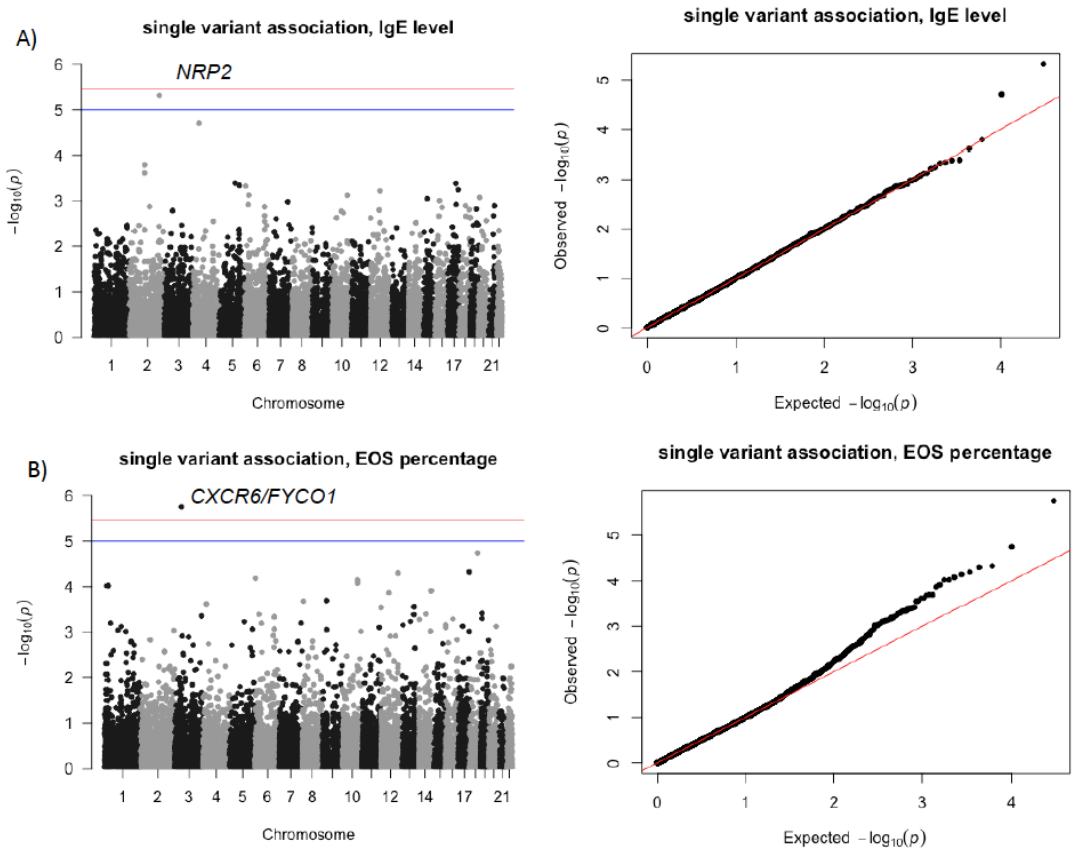
Supplementary Figure S8. Private low-frequency and singleton variants enrichment in A) Saguenay–Lac-Saint-Jean (SLSJ) and B) Finland (FINN) compared to France (FR), Sweden (SWE) and United Kingdom (UK).

Supplementary Figure S9



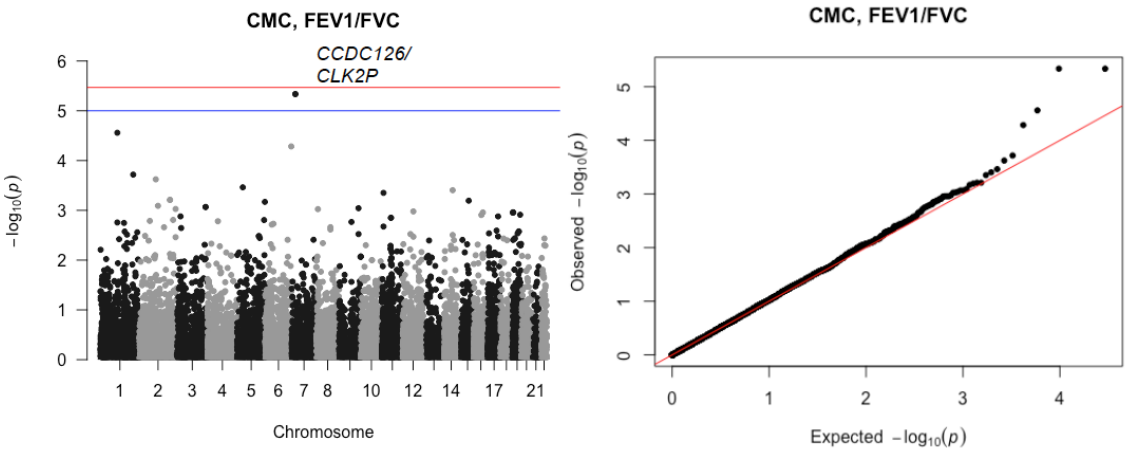
Supplementary Figure S9. GERP++ score distribution of private variants per population, A) All, B) non-synonymous, C) loss of Function, D) Synonymous. FINN: Finland, FR: France, SLSJ: Saguenay–Lac-Saint-Jean, SWE: Sweden, UK: United Kingdom.

Supplementary Figure S10



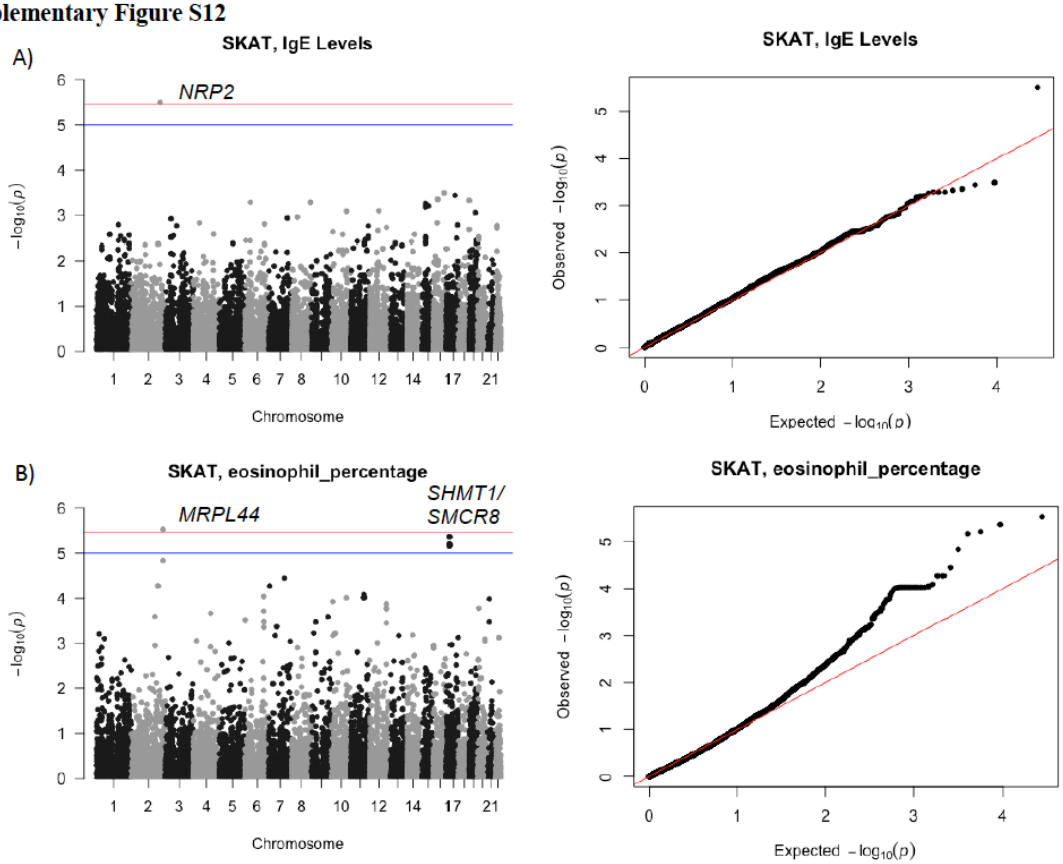
Supplementary Figure S10. Manhattan and qqplot for single variants association test for A) serum IgE Levels (Lambda = 1.009408) and B) Eos Percentage (Lambda= 1.075069).

Supplementary Figure S11



Supplementary Figure S11. Manhattan and qqplot for CMC test with FEV1/FVC (Lambda= 1.04)

Supplementary Figure S12



Supplementary Figure S12. Manhattan and qqplot for SKAT test with A) Serum IgE Levels (Lambda = 1.05) and B) Eos Percentage (Lambda= 1.13)

Chapter 4

Preface: Bridging Text between Chapters 3 and 4

Following on Chapter 3, we again used the SLSJ asthma familial cohort to explore the genetic basis of asthma and allergy-related traits. In this chapter, we used GWAS and DNA methylation data (EWAS) and linked them through mQTLs to identify new genes associated with allergic rhinitis with or without asthma.

Chapter 4: Combining omics data to identify genes associated with allergic rhinitis

Andréanne Morin[1,2], Michel Laviolette[3], Tomi Pastinen[1], Louis-Philippe Boulet[3] and Catherine Laprise[2,4]

[1] Department of Human Genetics, McGill University and Genome Quebec Innovation Centre, 740 Dr. Penfield Avenue, Montréal, Québec H3A 1A5 Canada

[2] Département des sciences fondamentales, Université du Québec à Chicoutimi, 555 boulevard de l'Université, Saguenay, Chicoutimi, Québec, G7H 2B1 Canada

[3] Institut Universitaire de Cardiologie et de Pneumologie de Québec, Université Laval, 2725, chemin Sainte-Foy, Québec, Québec, G1V 4G5 Canada

[4] Corresponding author

E-mail:

Andréanne Morin: andreanne.morin@mail.mcgill.ca

Michel Laviolette: michel.laviolette@med.ulaval.ca

Tomi Pastinen: tomi.pastinen@mcgill.ca

Louis-Philippe Boulet: lpboulet@med.ulaval.ca

Catherine Laprise: catherine.laprise@uqac.ca

Published in:
Clin Epigenetics. 2017 Jan 18;9:3. doi: 10.1186/s13148-017-0310-1.

4.1 Abstract

Allergic rhinitis is a common chronic disorder characterized by immunoglobulin E-mediated inflammation. To identify new genes associated with this trait, we performed genome- and epigenome-wide association studies and linked marginally significant CpGs located in genes or its promoter and SNPs located 1 Mb from the CpGs, by identifying *cis* methylation quantitative trait loci (mQTL). This approach relies on functional cellular aspects rather than stringent statistical correction. We were able to identify one gene with significant *cis*-mQTL for allergic rhinitis, caudal-type homeobox 1 (*CDX1*). We also identified 11 genes with marginally significant *cis*-mQTLs ($p < 0.05$) including one with both allergic rhinitis with or without asthma (*RNF39*). Moreover, most SNPs identified were not located closest to the gene they were linked to through *cis*-mQTLs counting the one linked to *CDX1* located in a gene previously associated with asthma and atopic dermatitis. By combining omics data, we were able to identify new genes associated with allergic rhinitis and better assess the genes linked to associated SNPs.

Key words: Allergic rhinitis, asthma, GWAS, EWAS, mQTLs, omics

4.2 Introduction

Allergic rhinitis is one of the most common allergy worldwide and one of the most common chronic disorders among children and adults [252]. Early sensitization to aeroallergens and food combined with the presence of atopic dermatitis, characterized by an immunoglobulin E (IgE)-mediated inflammation, can result in the development of asthma and/or allergic rhinitis later in life in a process called "atopic march" [87]. Genetic studies identified hundreds of genes associated with allergic rhinitis and genome-wide association studies (GWASs) pinpointed single nucleotide polymorphisms (SNPs) associated with its development [133, 135]. However, a majority of identified SNPs lie in the non-coding genomic region, making it difficult to identify the targeted genes. Given that DNA methylation may have an impact on gene regulation [253], the probability of detecting true positive associations should be improved by combining nominally significant data from genomics and epigenomics and linking them by quantitative trait loci (QTL) analysis. Methylation QTLs (mQTLs) allow assessing the impact of DNA sequenced variations (SNPs) on DNA methylation. They have been assessed in different tissues and cell types and were shown to overlap with GWAS hits [254-257]. We used this approach to identify allergic rhinitis genes and illustrate its usefulness in the context of a complex trait.

4.3 Materials and Methods

Individual selection, characterization, and sample preparation

We used data available from the Saguenay–Lac-Saint-Jean (SLSJ) asthma familial collection from Québec, Canada, that has data for rhinitis and allergies (Table 1). This population is know for its founder effect and is more homogeneous than a cosmopolitan population [258, 259]. Individuals affected with rhinitis and allergies, with or without asthma, were analyzed as cases. Individuals with no rhinitis, allergies and asthma were considered as controls. In this study, patients were defined as asthmatics based on if they either had a reported history of asthma (validated by a physician) or if at recruitment they manifested asthma-related symptoms and positive PC_{20} (<8 mg/ml) [186]. Rhinitis was self-reported and the subject had to answer "yes" to at least one of the following questions: Have you ever had rhinitis, Have you ever had hay fever, Have you ever had sneeze or rheum after a contact with: hay, flowers, animals, dust? Allergy was defined by a skin prick test for 26 aeroallergens (>=3mm). All subjects were recruited and evaluated out of the pollen season [186]. Recruitment and clinical evaluation of individuals as well as phenotype description can be found in Laprise 2014 [186]. All subjects gave their informed consent and the project was approved by the research ethic committee of the Centre intégré universitaire de santé et de services sociaux du SLSJ.

Genome-wide association study (GWAS)

A total of 508 subjects (321 cases and 187 controls) and 312 subjects (125 cases and 187 controls) were included in the analysis for allergic rhinitis with or without asthma respectively. The same group of control was used to compare to both phenotypes (i.e.

allergic rhinitis and allergic rhinitis with asthma). DNA extraction, genotyping methods and statistical analyses were described previously [186]. Genotyping was performed using the Illumina 610K Quad array (Illumina, San Diego, CA, USA). Association test was performed using a quasi-likelihood score test using the MQLS program (Release 1.5, http://www.stat.uchicago.edu/~mcpeek/software/MQLS/index.html), which allows performing case-control association analysis using related individuals [260]. The kinship coefficient was calculated using KinlnbCoef program (version 1.1, http://www.stat.uchicago.edu/~mcpeek/software/KinInbcoef/index.html). We included in the analysis SNPs with minor allele frequency (MAF) >0.05, p-value for Hardy Weinberg equilibrium >0.0001, and overall call rate >95%. Samples with genotyping rate <95% were excluded. A total of 633 samples (321 subjects with allergic rhinitis with asthma, 125 subject with allergic rhinitis only and 187 controls (used to compare to both phenotypes)) and 506,388 SNPs were included in the analysis.

Epigenome-wide association study (EWAS)

A total of 31 controls and 48 cases for allergic rhinitis with asthma or 30 cases for allergic rhinitis alone were included in the EWAS analysis. These samples are a subset of the ones used in the GWAS analysis. Unrelated subjects were included based on having allergic rhinitis with or without asthma, having no asthma, allergies or rhinitis, and based on having high or low levels of IgE. DNA extraction and sodium bisulfite conversion methods were described previously [197]. The assay was carried out on the Infinium HumanMethylation450 BeadChip array (Illumina, San Diego, CA, USA). The analysis was performed using the RnBeads Bioconductor R package [261]. We removed probes

with at least one of the following characteristics: (1) weak signal ($p>0.01$) (2,128 CpG sites), (2) SNP-enriched sites (4,100 sites), (3) out of a CpG context (not on a CG) (3,149 sites) or (4) located on sex chromosomes (11,129 sites). A total of 465,071 CpG sites were analyzed initially. Signal was then normalized, first by scaling to the internal controls using the methylumi R package [262], then applying the method of subset-quantile within array normalization (SWAN) implemented in the minfi R package [263, 264]. A total of 2,203 sites were removed due to missing data. We removed probes that mapped multiple genomic regions (≥90 % sequence similarity), have a variant less than 10bp from the CpG or that have ≥2 SNPs in it. A total of 374,498 CpG sites (80.5%) were analyzed for differential DNA methylation using limma package [265]. All samples had cell counts for eosinophils, basophils, monocytes, lymphocytes and neutrophils. The cell percentages were used as covariates as well as sex, age, smoking status, and batch effect.

Methylation quantitative trait loci analysis (mQTLs)

To perform the mQTL analyses, we used associated SNPs ($p<0.05$) and CpGs ($p<0.05$ and $\Delta\beta>0.05$) in the GWAS and EWAS for both traits. We kept associated CpGs that were located in either the gene body or 1.5kb upstream of the transcription start site, keeping 88 and 144 CpGs for allergic rhinitis with or without asthma respectively. SNPs were kept if present in all samples and if the three genotype groups (homozygous reference, heterozygous and homozygous alternative) were observed at least 5 times. A total of 529 and 625 SNPs were included in the analysis for allergic rhinitis with or without asthma respectively. We analyzed *cis*-mQTLs where the CpG-SNP combination

was less than 1Mb apart from each other based on the distance used by the GTEX consortia for their *cis*-eQTLs (http://www.gtexportal.org/home/documentationPage). We used a Bonferroni correction to evaluate significance thresholds. We computed mQTLs for these SNP-CpG pairs using an additive linear model using the R package MatrixEQTL [266]. Same covariates as in EWAS were included in this analysis. A total of 274 (Bonferroni p=0.05/274=1.8e-4) and 500 (Bonferroni p=0.05/500=1e-4) CpG-SNP comparisons were performed for allergic rhinitis with or without asthma respectively.

4.4 Results and Discussion

In this study, we used a novel approach that links genetics (SNPs) and functional (CpGs) data through the use of mQTLs identifying new genes associated to allergic rhinitis with or without asthma (Figure 1). It relies on functional cellular data and reduces the stringent cut-off normally used in GWAS. Even though this is a pilot experiment with small number of samples, we identified one significant *cis*-mQTL for allergic rhinitis located in *CDX1* (p=6.41e-5) (Table 2). We also observed nine nominally associated *cis*-mQTLs located in five genes for allergic rhinitis and 16 located in nine genes for allergic rhinitis with asthma (Table 2). One gene was reported being associated in both traits: *RNF39*. It has the highest number of mQTLs identified in both allergic rhinitis with (four) or without asthma (five).

The significantly or nominally associated genes were not associated to any related trait before. Interestingly, the majority of the genes linked to a SNP by the *cis*-mQTLs are not the closest ones, thus would not be the ones reported in a regular GWAS study. For example, all of the significant SNPs reported for the *RNF39 cis*-mQTLs are located 300Kb to 1Mb away from the gene and are located closer to other genes, which were previously associated with pulmonary function (rs2844833-*HLA-F* [267], rs2523872-*MUC22* [267], rs2517504-*HCG22* [154, 267], rs2535238-*ZFP57* [267]). The best example remains the one for the significantly associated mQTL that links rs888989 to a CpG located in the promoter region of the *CDX1* gene. The SNP is located in an intron of *TNIP1* and 900kb from *CDX1*. The former was previously associated to atopic dermatitis [137] and asthma [268]. According to the GTEx portal (http://www.gtexportal.org/), rs888989 and *CDX1* form an eQTL in the lungs (p=0.04), which is not the case for

TNIP1 (p=0.94). This reinforces the possible implication of this gene in allergic rhinitis and shows that our method may better assess the true genes of interest linked to the associated SNPs.

The originality of our approach resides in combining GWAS and EWAS nominally associated SNPs and CpGs, using *cis*-mQTL data, to identify genes of interest in this disease. This method has the potential to reduce false negative findings by relying on the cellular mechanisms of gene regulation compared to the use of stringent statistical corrections. The use of a well-described collection coming from a founder population and including subjects selected based on the same precise criteria allowed a more unified genetic background and phenotype. However, since this is a pilot study, the limited number of samples included in the EWAS and the GWAS may constraint the power of the findings. We were not able to test SNPs previously associated to the trait in previous GWASs because they did not meet the criteria to be included in the mQTL analysis. We also analyzed SNPs and CpGs preselected in the arrays by the manufacturers, thus excluding potentially important SNPs or CpG sites, which are not in linkage disequilibrium. DNA methylation analysis using whole blood could have limited the findings, even if correction for cell counts was included in our model. Apart from the limitations, we showed that our approach is promising and acknowledging for the lack of power in future studies will permit to better pinpoint genes of interests for different traits. Studying other tissues implicated in allergic rhinitis trait, like nasal or lung cells, could also reveal other genes implicated in the physiopathology. Genes identified in this study, notably *CDX1,* are worthwhile to be further investigated to understand the allergic rhinitis pathogenesis and the atopic march.

Declaration

Ethics approval and consent to participate

All subjects gave their informed consent and the project was approved by the research ethic committee of the Centre intégré universitaire de santé et de services sociaux du SLSJ.

Consent for publication

Not applicable

Availability of data and material

Data is available upon request

Competing interest

The authors declare no conflicts of interest

Funding

Canadian Institute of Health research operating grant

Authors' contribution

CL collected the data and managed the SLSJ cohort, conceived and supervised the study. AM analyzed, interpreted the data and wrote the manuscript draft under the supervision of CL. CL, LPB, ML and TP edited the manuscript. All authors reviewed and approved the final manuscript

Abbreviations

AR: Allergic rhinitis; ARA: Allergic rhinitis with asthma; CDX1: Caudal-type homeobox 1; eQTL: Expression quantitative trait loci; EWAS: Epigenomewide association study; GWAS: Genome-wide association study; HCG22: HLA complex group 22; HLA-F: Major histocompatibility complex, class I, F; IgE: Immunoglobulin E; MAF: Minor allele frequency; mQTL: Methylation quantitative trait loci; MUC22: Mucin 22; RNF39: Ring finger protein 39; SLSJ: Saguenay–Lac-Saint-Jean; SNP: Single nucleotide polymorphism; SWAN: Subset-quantile within array normalization; TNIP1: TNFAIP3 interacting protein 1; TSS: Transcription start site; ZFP57: Zinc finger protein 57

4.5 Acknowledgements

This work was supported by Laprise and Pastinen operating grants from the Canadian Institute of Health Research (CIHR), AM was supported by the Fonds de Recherche du Québec – Santé (FRQS) Doctoral training award. CL is the director of the Asthma Strategic Group of the Respiratory Health Network (RHN), investigator of CHILD Study and is a member of the AllerGen NCE Inc. CL is the chairholder of the Canada Research Chair in the Environment and Genetics of Respiratory Disorders and Allergies and TP is the chairholder of the Canada Research Chair in Human Genomics.

4.6 Figure and Tables

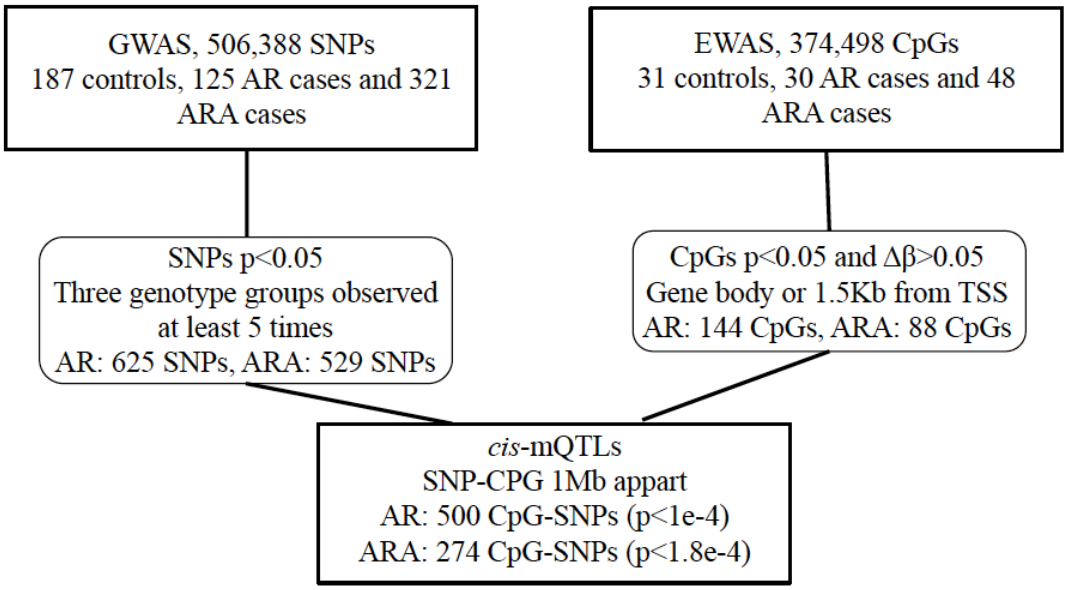


Figure 1. Flowchart presenting our approach combining genome-wide association study (GWAS) and epigenome-wide association study (EWAS) hits to identify *cis* methylation quantitative trait loci (mQTLs) that could be association to allergic rhinitis with (ARA) or without asthma (AR). We first performed GWAS and EWAS separately for AR and ARA. We then selected marginally associated SNPs ($p<0.05$) where the three genotyping groups were observed at least five times. We also selected marginally associated CpGs ($p<0.05$) that had a $\Delta\beta>0.05$ and that were located in the gene body or 1.5Kb from the transcription start site (TSS). We then linked the SNPs and CpGs that were 1Mb apart by performing *cis*-mQTLs for both AR and ARA. We used Bonferonni p-value cut-offs to assess significance.

Table 1. General characterization of individuals analyzed in the study

	GWAS samples			EWAS samples		
	Controls[a]	**Allergic rhinitis[b]**	**Allergic rhinitis combined with asthma[c]**	**Controls[a]**	**Allergic rhinitis[b]**	**Allergic rhinitis combined with asthma[c]**
Number of samples	187	125	321	31	30	48
M:F ratio	1:1.13	1:0.87	1:0.87	1:1.60	1:0.88	1:0.78
Age, mean (range)[d]	43 (3-85)	37 (5-93)	28 (5-83)	29 (1-53)	28 (1-59)	28 (5-55)
Age median[d]	41	38	26	35	30	26
Smoking status, n (%)[e]						
Non smoker	82 (44)	64 (51)	219 (68)	14 (45)	18 (60)	36 (75)
Ex smoker	61 (33)	37 (30)	53 (17)	8 (26)	6 (20)	4 (8)
Smoker	43 (23)	21 (17)	44 (14)	9 (29)	5 (17)	7 (15)
IgE, (SD)[f]	202.85 (1373.66)	411.27 (852.17)	856.45 (2075.62)	67.10 (90.45)	575.40 (1380.45)	597.73 (242.50)

[a] Defined as not affected by either asthma, allergies or rhinitis. [b] Defined as being affected with both allergy and rhinitis. Allergic rhinitis phenotype is available for all samples. Allergy is defined as at least one positive response on skin prick testing (wheal diameter ≥ 3 mm at 10 minutes). Rhinitis is self-reported and the subject had to answer "yes" to at least one of the following questions: Have you ever had rhinitis, Have you ever had hay fever, Have you ever had sneeze or rheum after a contact with: hay, flowers, animals, dust? Can be either [c] combined or [b] not with asthma. [d] Age difference between groups were assess using an unpaired t-test. GWAS: controls vs allergic rhinitis p=0.078, control vs Allergic rhinitis combined with asthma p=1.2e-15. EWAS: controls vs allergic rhinitis p=0.078, control vs Allergic rhinitis combined with asthma p=0.43. [e] Smoking status available for 186 controls, 122 allergic rhinitis and 316 allergic rhinitis combined with asthma subjects for genome-wide association study (GWAS) samples, 31 controls, 29 allergic rhinitis and 47 allergic rhinitis combined with asthma subjects for epigenome-wide association study (EWAS) samples. Differences between groups were assessed using a chi-square test. GWAS: controls vs allergic rhinitis p=0.0045, control vs Allergic rhinitis combined with asthma p=1.25e-19. EWAS: controls vs allergic rhinitis p=0.049, control vs Allergic rhinitis combined with asthma p=7.7e-3. [f] Geometric mean and standard deviation (SD) for the Immunoglobulin E (IgE) serum concentration calculated for 175 controls, 116 allergic rhinitis and 302 allergic rhinitis combined with asthma subjects for GWAS samples and all subjects for EWAS samples. IgE levels difference between groups were assess using an unpaired t-test. GWAS: controls vs allergic rhinitis p=0.145, control vs Allergic rhinitis combined with asthma p=2.2e-3. EWAS: controls vs allergic rhinitis p=0.003, control vs Allergic rhinitis combined with asthma p=0.90. Sex, age, cell count and smoking status were used as covariates in the analysis.

Table 2. Genes with *cis*-mQTL sites significantly associated with allergic rhinitis with or without asthma.

			mQTLs	GWAS analysis		EWAS analysis		
Trait	Gene	Locus	p-value	SNP	p-value	CpGs	Δβ[a]	p-value
Allergic rhinitis	***CDX1***	chr5q32	**6.41e-5**	rs888989	**0.0038**	cg18424208	-5.19	**0.0002**
	PPAN-P2RY11	chr19p13.2	0.0245	rs3752199	0.0346	cg24118856	7.51	4.39e-5
			0.0090	rs2844833	0.0270	cg05563515	10.11	0.0212
			0.0229	rs2844833	0.0270	cg24637044	5.85	0.0132
	*RNF39**	chr6p22.1	0.0265	rs2844833	0.0270	cg01286685	7.78	0.0266
			0.0411	rs2523872	0.0123	cg10930308	9.50	0.0255
			0.0499	rs2523872	0.0123	cg01286685	7.78	0.0266
	SRRT	chr7q22.1	0.0412	rs6942824	0.0224	cg10426581	5.26	0.0096
Allergic rhinitis with asthma	*ADORA1*	chr1q32.1	0.0337	rs6661284	0.0337	cg19315653	-6.26	0.0315
	ITGB2	chr21q22.3	0.0381	rs7275203	0.0381	cg18012089	6.10	0.0068
	LINC00336	chr6p21.31	0.0073	rs9461924	0.0073	cg04329454	-7.16	0.0015
	MFSD6L	chr17p13.1	0.0120	rs9895992	0.0120	cg11685316	5.01	0.0072
	PCDH8	chr13q14.3	0.0152	rs732774	0.0295	cg14950829	7.53	0.0097
			0.0135	rs3742297	0.0480	cg14950829	7.53	0.0097
			0.0259	rs1801249	0.0296	cg14950829	7.53	0.0097
			0.0259	rs4943046	0.0298	cg14950829	7.53	0.0097
	PITX2	chr4q25	0.0257	rs2067004	0.0272	cg13385016	5.06	0.0240
			0.0249	rs9992755	0.0289	cg13385016	5.06	0.0240
	RNF180	chr5q12.3	0.0130	rs7713289	0.0130	cg17370163	5.43	0.0021
	*RNF39**	chr6p22.1	0.0133	rs2517504	0.0047	cg03343571	9.19	0.0451
			0.0171	rs2517504	0.0047	cg01286685	8.21	0.0478
			0.0401	rs2535238	0.0248	cg01286685	8.21	0.0478
			0.0499	rs2523872	0.0299	cg01286685	8.21	0.0478
	ZFPM1	chr16q24.2	0.0304	rs750740	0.0304	cg04983687	5.53	0.0056

[a] Δβ and p-values for CpG sites and SNPs forming a *cis*-mQTL. A negative Δβ indicates a decrease in the percentage of methylation for cases compared to controls. All locus refer to the human hg19 reference genome. * *RNF39* is the only gene marginally associated in both traits.

Chapter 5: Discussion and future directions

The purpose of this work was to go beyond GWAS studies in order to better understand the genetic basis of complex traits, using asthma and allergic diseases as an example. GWAS were highly important in understanding the different genetic architecture of complex traits, confirming previously identified genes and uncover new ones. The high expectation of these studies led to their increasing popularity in the mid 2000s and brought a tremendous amount of knowledge that helped better understand complex traits. However, one drawback is that even large-scale studies including thousands of individuals do not explain the whole picture [14]. This ascertainment led the research community to develop new strategies to complement GWAS limitations. The work presented in this thesis and as part of my PhD degree aims to explore different strategies to better understand the genetic and epigenetic bases of asthma and allergy related-traits. Specifically, we decided to explore two different strategies to study asthma and allergy related-traits: investigating rare and low-frequency variants and linking GWAS hits to cellular traits (DNA methylation and gene expression). As mentioned in Chapter 1, asthma and allergies affect a large number of individuals, especially in developed countries. It results in more than 250,000 deaths per year and represents an important economic burden [85]. A lot of effort has been invested in studying these diseases despite the fact that they are very difficult to investigate. They are highly heterogeneous, being clinically modulated by environmental and genetic determinants. They are also seen as a plethora of different diseases that are sometimes hard to differentiate from one another. One of the strengths of the studies presented herein is the use of the SLSJ cohort, a founder population for which very detailed phenotypic information is available.

Rare and low-frequency variants in complex traits

The Chapter 2 of this thesis describes a custom capture panel designed to study rare and low-frequency variants in autoimmune and inflammatory diseases in a cost-effective manner. Based on previous knowledge acquired from GWAS, we wanted to assess the contribution of coding and non-coding regions in such diseases. We selected interesting non-coding functional regions to study based on whole-genome DHS mapping from different immune cells, which have been shown to be enriched in GWAS hits [17]. We designed the panel to study relevant cell-types involved in multiple immune and inflammatory diseases with the objective of later applying it to explore their genetic basis. We showed that the variants captured were highly functional and had an impact on gene expression. Using high-throughput next-generation sequencing allowed us to uncover new variants that could not have been identified using other technologies like classical genotyping chip or Immunochip. It also permits exploring non-coding regulatory regions as compared to whole-exome sequencing and is more cost effective than whole-genome sequencing, thus allowing us to sequence more individuals at a deeper coverage.

In Chapter 3, we showed that our custom capture panel was successfully used in identifying new genes and variants associated with asthma and allergy related-traits. To this end, we used the SLSJ asthma familial cohort, which has three main advantages: 1) it is a founder population, 2) it includes large families and 3) samples have high quality phenotype characterization (testing was done for all participants and phenotypes were not self-reported). The two first characteristics were important to limit the amount of genetic heterogeneity, which is an important obstacle in rare and low-frequency variants studies [65]. The founder

population allowed to enrich for higher frequency deleterious and private variants that could not be tested using other populations as shown in Chapter 3. They have also been successful in identifying rare variants associated with complex traits in the past [68, 183-185]. Family studies can also be an asset in the study of rare and low-frequency variants in complex trait because predisposing variants can be observed at a much higher frequency in affected members of the family, since multiple affected members may carry the same variant. It can also help reducing population structure bias by decreasing heterogeneity and achieve discoveries with a smaller sample size [65]. Future studies using this data could also explore *de novo* mutations or parent-of-origin effect that could be implicated in the development of the disease.

The third strength of this study was the fact that all samples were well characterized, having access quantitative disease-related traits like serum IgE levels, eosinophil counts and percentage as well as lung function measurements. These subphenotypes helped us to get enough power to observe association in a smaller sample set. We were able to identify significantly associated low-frequency variants with eosinophil percentage and located in two genes: *CXCR6* and *FYCO1*. We also observed two significant genes in collapsing analyses: *MRPL44* associated with eosinophil percentage and *NRP2* associated with serum IgE levels. Some variants showed increased frequency compared to previously assessed European populations, which allowed us identifying some variants associated with the traits and located in or close to genes never identified before that could help better understand disease biology. In fact, variants taking part into the autophagy process or chemokine receptors were identified, supporting the importance of these pathways that were previously implicated in the pathophysiology of these diseases [244, 269], but more

interestingly in the SLSJ asthma familial cohort [189, 191, 270, 271]. However, the identification of these variants could result in difficulties of replication and highlights the importance of population specific studies. The two variants identified in this study using single variant association test were non-coding as well as two of the lead SNV identified in the gene-based test. These variants would not have been identified using whole-exome sequencing or genotyping chip as they were also rare or low-frequency. These results underline the interesting aspect of using our custom capture panel on the SLSJ asthma familial cohort.

In addition to the results obtained on asthma and allergy-related traits, ongoing work uses our Immune-genetics sequencing to study other autoimmune and inflammatory diseases. The lower cost of our method allowed to sequence coding and regulatory non-coding regions of over 5000 samples from individuals affected with multiple diseases such as Multiple Sclerosis, Crohn's disease, Systemic Lupus Erythematosus, etc. We will try to identify genes that are shared across diseases (pleiotropic effect) as well as the disease-specific ones to better understand the genetic basis of these diseases.

Linking cellular traits to GWAS hits

So far, studies linking genetic variation to cellular traits use different strategies. Two of them are either to 1) link significant GWAS hits to functional traits like gene expression or DNA methylation or 2) combine directly the GWAS data to functional data focusing also on marginally associated sites. We used the latter one in our study, showing that we can achieve association with a smaller sample set using this strategy. We showed that the use of cellular and functional traits could help separating the true signals from the noise.

The Chapter 4 of this thesis describes a strategy to combine different omics data to identify new genes associated with allergic rhinitis with or without asthma. We took again advantage of the SLSJ asthma familial cohort. For this study, the population was a great advantage being more homogeneous than cosmopolitan populations, not only at the genetic level (families and founder effect) but also at the environmental one. Using families from a specific region allowed us to have samples sharing very similar environment, life habits, religion, diet, etc., thus reducing its potential effect on DNA methylation. Another strength from this study is the well-described samples. We used stringent phenotype inclusion criteria for allergic rhinitis with or without asthma. The samples were also evaluated clinically using a defined protocol by Dr. Laprise and the diagnoses confirmed by the same group of physicians (Dr. Bégin for adults and Dr. Morin for children).

We combined marginally associated SNPs and CpGs from GWAS and EWAS studies for allergic rhinitis with or without asthma. This could help departing the SNPs that do not resist correction in GWAS from false negative sites, thus being still true associations. In this study, we were able to identify a statistically significant mQTL of a CpG located in the promoter region of the *CDX1* gene linked to a SNP located in the intron region of a gene located 900KB from it. We not only identified a novel gene potentially associated with the trait, we also showed that the marginally associated SNP had an impact on a much further gene rather than the closest one. We showed here, like many others, which was done in many GWAS study, is not necessarily the proper way to do it. The SNP taking part in the significant mQTL is located in the intron of *TNIP1* gene which was previously associated with asthma and atopic dermatitis [137, 268].

Chapter 2 of this thesis showed the potential impact of rare variants on gene expression in T cells. It was one of the first studies to look at the impact of rare variants on gene expression using ASE. Although all studies used different ways to assess the functional impact of rare variants showing the difficult aspect of it, they do appear to take part in regulating gene expression. Our study adds to the body of literature highlighting the difficulty to explore this question, but also the importance of considering their implication in the genetic architecture of gene expression [48-50]. They could help in the future to better understand the functional aspect of rare and low-frequency non-coding variants associated with complex traits.

In Chapter 3, we used epigenetic data to support our finding of rare and low-frequency variants associated with asthma and allergy-related traits. We observed an enrichment of marginally significantly associated CpGs ($p<0.05$) located in the vicinity of the *NRP2* genes. These results were obtained using DNA methylation data in isolated eosinophils, but were not observed in whole-blood. These results support the importance of using specific cell-type that plays an important role in the pathophysiology and their consideration in future studies.

Linking SNPs and rare variants to gene expression and DNA methylation alone can help understand the functional aspect of genetic variants. However, using them individually only gives part of the explanation. Linking variants to gene expression can identify the target genes and if the allele reduces or increases the expression of the genes. This is actually very useful since a lot of the first GWAS studies assumed that the affected genes were the ones located closest to the SNPs. However, it does not give insight on how the expression is actually regulated. The latter can be explained by linking genetic variants to DNA methylation where the specific region or sequence elements can be pinpointed

(promoter, enhancer, gene body or insulator). But it does not always link the variants to the gene of interest when it is located away from it. Future studies exploring multiple epigenetics layers could help better understand genetic loci associated with diseases.

As an example, in a study performed by two other postdoctoral fellows in the lab and myself [34] we assessed the genetic effect on DNA methylation, histone deposition and gene expression. We also observed their interrelation at a high resolution. We performed allelic and non-allelic correlation between gene expression and DNA methylation and observed a higher rate of strong correlation using allele specific assessments. When combining ASM, ASE and allele-specific histone deposition (ASH), we saw a high concordance between high gene expression and high chromatin modification rate with active enhancer marks (H3K27ac, H3K4me1 and H3K4me3) when the linked CpG harbored lower methylation. The opposite was observed for repressive marks (H3K27me3, H3K36me3, H3K9me3). The effect was stronger when focusing on significant sites ($p<0.05$). These results highlight the sensitive detection that allele-specific analyses can bring to reveal links between multiple layers of functional features. These results also show the potential of using multiple layers to understand at a deeper level the functional impact of associated variants. In this study, we also observed that over 50% of mQTLs and over 25% of ASM identified were cell-type specific. We also observed an enrichment of autoimmune disease GWAS hits for ASM in naïve T cells and at a lower extent in ASM from whole-blood highlighting the importance of using isolated cells and tissues to gain more sensitivity to identify variants and understand their functional impact. Thus, combining tissue-specific DNA methylation and gene

expression data could provide a much deeper understanding of autoimmune diseases compared to genetic data alone.

Using better phenotyping

In the recent years, asthma has been increasingly seen as a combination of multiple phenotypes rather than a single disease. In fact, asthma is a combination of numerous clinical and physiological features that have been used to differentiate the diverse endotypes [272]. They will become more and more precise as the amount of data available increases, thus will probably take part into precision medicine where disease prevention, intervention and treatment will be fitted to each patient.

The first way to differentiate different asthma endotypes was based on the presence or not of allergic features as well as age of onset [272]. Early-onset asthmatics had mainly atopic and allergic triggers in combination with other allergic disease like rhinitis or atopic dermatitis. Asthmatics individuals that developed the disease later in life did not have allergic sensitization linked to it. More recently biological and genomic features were included to better define them [272]. One of the mostly known is the T_H2 process that is linked to allergy, atopy and eosinophilic inflammation. T_H2 associated asthma is also known to be corticosteroid responsive. This feature is a characteristic of early-onset allergic, late-onset eosinophilic as well as exercise induced asthma. The well-known 17q21 loci was also mainly associated with early-onset asthma but not to atopy or adult-onset asthma [121, 273]. Other endotypes do not display T_H2 characteristics like obesity-related and neutrophilic asthma, which are poor corticosteroids responders. Following patients' characterization using T_H2 features, nature of inflammation (e.g. eosinophilic versus neutrophilic) and treatment response (e.g. steroid responsive versus steroid

resistant), the advent of "omics" data helped decipher underlying mechanisms leading to more specific characterization of the disease. An example of that are the severe asthma endotypes proposed by Poon *et al.* characterized by different cytokines pathways [274]. For example, the IL-4/IL-13 pathway was linked to severe asthmatics with high IgE levels, the IL-5/IL-33 was linked to inhaled corticosteroid poor responders and IL-17 with neutrophilic asthma. The endotypes were also a great tool to identify new genes associated with the traits. For example, a study by Bønnelykke *et al.* performed a GWAS for early-onset asthma patients characterized by recurrent and severe exacerbations [275]. They identified previously known loci (*GSDMB*, *IL33*, *RAD50* and *IL1RL1*) at an effect size much larger than other studies and a new susceptibility loci underlining the strength of using specific phenotypes in the search for genes associated with the trait.

In this thesis, we tried to use better phenotyping data to decipher the genetics of asthma and allergic diseases. In fact, by its ascertainment, the SLSJ cohort is mainly composed of T_H2 associated asthmatics individuals. In chapter 3, we used phenotypes that are specific to allergic asthma like serum IgE levels and eosinophil percentage, which were the two phenotypes where we could identify significant genes and variants. We also used lung function measures for which a reduction was previously more linked to severe asthmatics [276], which are less present in the SLSJ familial cohort. In chapter 4 we assessed the genetic background of individuals affected with allergic rhinitis with and without asthma, which is linked to T_H2 response type. We can thus state that we took advantage of refined phenotyping data and encourage the use of endotypes to be able to identify new genes associated with asthma and allergy-related traits.

In the future, studies will not only have to include refined phenotyping data and endotypes to explore the genetic background of asthma and allergy-related traits, but also data about environmental exposure and known genotyping background. In fact, studies showed that some associations were only present in subjects exposed to certain environment. For example, an association with the 17q21 loci was only present in children who wheezed and were exposed to rhinovirus infection in the first three years of life [277, 278]. Taking into account environmental exposure allowed to observe a much higher odd ratio of 6.9 compared to 1.2 in other large consortia where they were not taken into consideration [121, 279]. Another example of the importance of including environmental exposures as well as genetic background was observed when looking at the effect of animal shed exposure on asthma development. Loss *et al.* observed that the protective effect of the exposure was genotype-dependent: one allele conferred protective effect for asthma and wheeze when exposed to animal shed, whereas the protective effect of the exposure was not observed in the presence of the other allele [278]. Another interesting finding was that the protective effect of the allele was only observed in the combination with the environmental exposure; no protective effect was observed when the children were not exposed to animal sheds [278]. These examples show the importance and the potential benefits of incorporating environmental and genetic background in future studies, in addition to multiple epigenetic layers and cellular traits as described in this thesis.

Conclusion

The findings presented in this thesis represent contribution in understanding the genetic basis of complex traits in multiple ways. We first developed a custom capture panel that allowed us to explore both coding and non-coding regulatory rare variants in a cost-effective manner. We showed the potential impact of rare variants on gene expression as well as we explored their effect along with low-frequency ones in asthma and allergy related traits. We uncovered three genes that were not identified before, but were part of processes like autophagy or chemokine receptors that have been implicated in the traits. We showed the potential of our target capture and how it can be used to explore the contribution of rare and low-frequency variants in other immune related diseases in the future. We used DNA methylation to confirm our findings in the rare variants analysis, but also to identify new genes associated with allergic rhinitis. In this thesis, we showed the potential of going beyond GWAS findings, learning lessons and complementing its limitations to better understand the genetic of complex trait such as asthma and allergy-related traits.

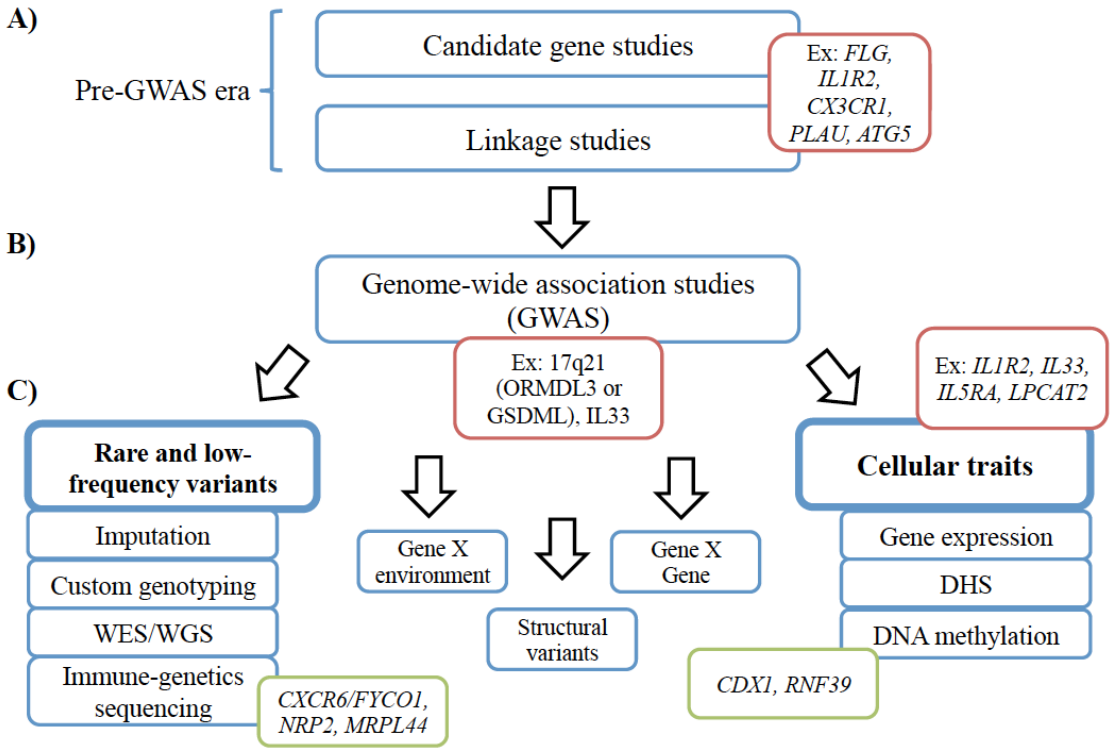


Figure 1. Representation of pre and post-GWAS era genetic approaches to study complex traits. A) Candidate genes and linkage studies were one of the first ways to study the genetic aspect of complex traits. A few genes were identified in the SLSJ asthma cohort using one of the techniques and examples are listed in the red boxes. B) These studies were followed by the advent of Genome-wide association studies (GWAS). The SLSJ asthma familial collection took part of large consortium that identified loci the were highly replicated. Examples of these loci are listed in the red box. C) New strategies were developed to complement GWAs findings. Two of them (in bold) are explored in this thesis: assessment of rare and low-frequency variants and linking GWAs hits to cellular traits. Different ways to explore these two strategies are also listed as well as examples of genes that were identified in the SLSJ asthma familial cohort (red boxes). In the green boxes are listed the additional genes identified using these post-GWAS era strategies that were identified in this thesis.

References

1. Altmuller, J., et al., *Genomewide scans of complex human diseases: true linkage is hard to find.* Am J Hum Genet, 2001. **69**(5): p. 936-50.
2. Hirschhorn, J.N., et al., *A comprehensive review of genetic association studies.* Genet Med, 2002. **4**(2): p. 45-61.
3. Todd, J.A., *Statistical false positive or true disease pathway?* Nat Genet, 2006. **38**(7): p. 731-3.
4. Lander, E.S., et al., *Initial sequencing and analysis of the human genome.* Nature, 2001. **409**(6822): p. 860-921.
5. Venter, J.C., et al., *The sequence of the human genome.* Science, 2001. **291**(5507): p. 1304-51.
6. International HapMap, C., *The International HapMap Project.* Nature, 2003. **426**(6968): p. 789-96.
7. Klein, R.J., et al., *Complement factor H polymorphism in age-related macular degeneration.* Science, 2005. **308**(5720): p. 385-9.
8. Hindorff LA, M.J.E.B.I., Morales J (European Bioinformatics Institute), Junkins HA, Hall PN, Klemm AK, and Manolio TA. *A Catalog of Published Genome-Wide Association Studies. Available at:* http://www.genome.gov/gwastudies. *Accessed 29-01-2015.*
9. Agarwala, V., et al., *Evaluating empirical bounds on complex disease genetic architecture.* Nat Genet, 2013. **45**(12): p. 1418-27.
10. Voight, B.F., et al., *Plasma HDL cholesterol and risk of myocardial infarction: a mendelian randomisation study.* Lancet, 2012. **380**(9841): p. 572-80.
11. Spycher, B.D., et al., *Genome-wide prediction of childhood asthma and related phenotypes in a longitudinal birth cohort.* J Allergy Clin Immunol, 2012. **130**(2): p. 503-9 e7.
12. Xu, M., et al., *Genome Wide Association Study to predict severe asthma exacerbations in children using random forests classifiers.* BMC Med Genet, 2011. **12**: p. 90.
13. Sanseau, P., et al., *Use of genome-wide association studies for drug repositioning.* Nat Biotechnol, 2012. **30**(4): p. 317-20.
14. Manolio, T.A., et al., *Finding the missing heritability of complex diseases.* Nature, 2009. **461**(7265): p. 747-53.
15. Welter, D., et al., *The NHGRI GWAS Catalog, a curated resource of SNP-trait associations.* Nucleic Acids Res, 2014. **42**(Database issue): p. D1001-6.
16. Williams, S.M. and J.L. Haines, *Correcting away the hidden heritability.* Ann Hum Genet, 2011. **75**(3): p. 348-50.
17. Maurano, M.T., et al., *Systematic localization of common disease-associated variation in regulatory DNA.* Science, 2012. **337**(6099): p. 1190-5.
18. Nicolae, D.L., et al., *Trait-associated SNPs are more likely to be eQTLs: annotation to enhance discovery from GWAS.* PLoS Genet, 2010. **6**(4): p. e1000888.
19. Wei, W.H., G. Hemani, and C.S. Haley, *Detecting epistasis in human complex traits.* Nat Rev Genet, 2014. **15**(11): p. 722-33.

20. Thomas, D., *Gene--environment-wide association studies: emerging approaches.* Nat Rev Genet, 2010. **11**(4): p. 259-72.
21. Mefford, H.C. and E.E. Eichler, *Duplication hotspots, rare genomic disorders, and common disease.* Curr Opin Genet Dev, 2009. **19**(3): p. 196-204.
22. Visscher, P.M., W.G. Hill, and N.R. Wray, *Heritability in the genomics era--concepts and misconceptions.* Nat Rev Genet, 2008. **9**(4): p. 255-66.
23. Golan, D., E.S. Lander, and S. Rosset, *Measuring missing heritability: inferring the contribution of common variants.* Proc Natl Acad Sci U S A, 2014. **111**(49): p. E5272-81.
24. Yang, J., et al., *Common SNPs explain a large proportion of the heritability for human height.* Nat Genet, 2010. **42**(7): p. 565-9.
25. Speed, D., et al., *Re-evaluation of SNP heritability in complex human traits.* bioRxiv, 2017.
26. Fuchsberger, C., et al., *The genetic architecture of type 2 diabetes.* Nature, 2016. **536**(7614): p. 41-7.
27. Igartua, C., et al., *Ethnic-specific associations of rare and low-frequency DNA sequence variants with asthma.* Nat Commun, 2015. **6**: p. 5965.
28. Hunt, K.A., et al., *Negligible impact of rare autoimmune-locus coding-region variants on missing heritability.* Nature, 2013. **498**(7453): p. 232-5.
29. Nelson, M.R., et al., *An abundance of rare functional variants in 202 drug target genes sequenced in 14,002 people.* Science, 2012. **337**(6090): p. 100-4.
30. Marouli, E., et al., *Rare and low-frequency coding variants alter human adult height.* Nature, 2017. **542**(7640): p. 186-190.
31. Kong, A., et al., *Parental origin of sequence variants associated with complex diseases.* Nature, 2009. **462**(7275): p. 868-74.
32. Nadeau, J.H., *Transgenerational genetic effects on phenotypic variation and disease risk.* Hum Mol Genet, 2009. **18**(R2): p. R202-10.
33. Pastinen, T. and T.J. Hudson, *Cis-acting regulatory variation in the human genome.* Science, 2004. **306**(5696): p. 647-50.
34. Cheung, W.A., et al., *Functional variation in allelic methylomes underscores a strong genetic contribution and reveals novel epigenetic alterations in the human epigenome.* Genome Biol, 2017. **18**(1): p. 50.
35. Gross, D.S. and W.T. Garrard, *Nuclease hypersensitive sites in chromatin.* Annu Rev Biochem, 1988. **57**: p. 159-97.
36. Thurman, R.E., et al., *The accessible chromatin landscape of the human genome.* Nature, 2012. **489**(7414): p. 75-82.
37. Elangovan, R.I., et al., *Regulatory genomic regions active in immune cell types explain a large proportion of the genetic risk of multiple sclerosis.* J Hum Genet, 2014. **59**(4): p. 211-5.
38. Mokry, M., et al., *Extensive Association of Common Disease Variants with Regulatory Sequence.* PLoS One, 2016. **11**(11): p. e0165893.
39. Pastinen, T., *Genome-wide allele-specific analysis: insights into regulatory variation.* Nat Rev Genet, 2010. **11**(8): p. 533-8.
40. Ge, B., et al., *Global patterns of cis variation in human cells revealed by high-density allelic expression analysis.* Nat Genet, 2009. **41**(11): p. 1216-22.

41. Almlof, J.C., et al., *Powerful identification of cis-regulatory SNPs in human primary monocytes using allele-specific gene expression.* PLoS One, 2012. **7**(12): p. e52260.
42. Dimas, A.S., et al., *Common regulatory variation impacts gene expression in a cell type-dependent manner.* Science, 2009. **325**(5945): p. 1246-50.
43. Consortium, G.T., *The Genotype-Tissue Expression (GTEx) project.* Nat Genet, 2013. **45**(6): p. 580-5.
44. Grundberg, E., et al., *Global analysis of DNA methylation variation in adipose tissue from twins reveals links to disease-associated variants in distal regulatory elements.* Am J Hum Genet, 2013. **93**(5): p. 876-90.
45. Hutchinson, J.N., et al., *Allele-specific methylation occurs at genetic variants associated with complex disease.* PLoS One, 2014. **9**(6): p. e98464.
46. Chuang, L.C., et al., *Pathway analysis using information from allele-specific gene methylation in genome-wide association studies for bipolar disorder.* PLoS One, 2013. **8**(1): p. e53092.
47. Do, C., et al., *Mechanisms and Disease Associations of Haplotype-Dependent Allele-Specific DNA Methylation.* Am J Hum Genet, 2016. **98**(5): p. 934-55.
48. Li, X., et al., *Transcriptome sequencing of a large human family identifies the impact of rare noncoding variants.* Am J Hum Genet, 2014. **95**(3): p. 245-56.
49. Zeng, Y., et al., *Aberrant gene expression in humans.* PLoS Genet, 2015. **11**(1): p. e1004942.
50. Zhao, J., et al., *A Burden of Rare Variants Associated with Extremes of Gene Expression in Human Peripheral Blood.* Am J Hum Genet, 2016. **98**(2): p. 299-309.
51. Pala, M., et al., *Population- and individual-specific regulatory variation in Sardinia.* Nat Genet, 2017. **49**(5): p. 700-707.
52. Richardson, T.G., et al., *Collapsed methylation quantitative trait loci analysis for low frequency and rare variants.* Hum Mol Genet, 2016. **25**(19): p. 4339-4349.
53. Zuk, O., et al., *Searching for missing heritability: designing rare variant association studies.* Proc Natl Acad Sci U S A, 2014. **111**(4): p. E455-64.
54. Cortes, A. and M.A. Brown, *Promise and pitfalls of the Immunochip.* Arthritis Res Ther, 2011. **13**(1): p. 101.
55. Genomes Project, C., et al., *A map of human genome variation from population-scale sequencing.* Nature, 2010. **467**(7319): p. 1061-73.
56. 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.
57. 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.
58. Eyre, S., et al., *High-density genetic mapping identifies new susceptibility loci for rheumatoid arthritis.* Nat Genet, 2012. **44**(12): p. 1336-40.
59. 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.

60. Liu, J.Z., et al., *Association analyses identify 38 susceptibility loci for inflammatory bowel disease and highlight shared genetic risk across populations.* Nat Genet, 2015. **47**(9): p. 979-86.
61. Voight, B.F., et al., *The metabochip, a custom genotyping array for genetic studies of metabolic, cardiovascular, and anthropometric traits.* PLoS Genet, 2012. **8**(8): p. e1002793.
62. Grove, M.L., et al., *Best practices and joint calling of the HumanExome BeadChip: the CHARGE Consortium.* PLoS One, 2013. **8**(7): p. e68095.
63. Genomes Project, C., et al., *An integrated map of genetic variation from 1,092 human genomes.* Nature, 2012. **491**(7422): p. 56-65.
64. Tennessen, J.A., et al., *Evolution and functional impact of rare coding variation from deep sequencing of human exomes.* Science, 2012. **337**(6090): p. 64-9.
65. Consortium, U.K., et al., *The UK10K project identifies rare variants in health and disease.* Nature, 2015. **526**(7571): p. 82-90.
66. Park, J.H., et al., *Distribution of allele frequencies and effect sizes and their interrelationships for common genetic susceptibility variants.* Proc Natl Acad Sci U S A, 2011. **108**(44): p. 18026-31.
67. Smith, D., et al., *A rare IL33 loss-of-function mutation reduces blood eosinophil counts and protects from asthma.* PLoS Genet, 2017. **13**(3): p. e1006659.
68. Sidore, C., et al., *Genome sequencing elucidates Sardinian genetic architecture and augments association analyses for lipid and blood inflammatory markers.* Nat Genet, 2015. **47**(11): p. 1272-81.
69. Genome of the Netherlands, C., *Whole-genome sequence variation, population structure and demographic history of the Dutch population.* Nat Genet, 2014. **46**(8): p. 818-25.
70. International HapMap, C., *A haplotype map of the human genome.* Nature, 2005. **437**(7063): p. 1299-320.
71. International HapMap, C., et al., *A second generation human haplotype map of over 3.1 million SNPs.* Nature, 2007. **449**(7164): p. 851-61.
72. Huang, J., et al., *Improved imputation of low-frequency and rare variants using the UK10K haplotype reference panel.* Nat Commun, 2015. **6**: p. 8111.
73. Pasaniuc, B., et al., *Extremely low-coverage sequencing and imputation increases power for genome-wide association studies.* Nat Genet, 2012. **44**(6): p. 631-5.
74. Zheng, H.F., et al., *Effect of genome-wide genotyping and reference panels on rare variants imputation.* J Genet Genomics, 2012. **39**(10): p. 545-50.
75. McCarthy, S., et al., *A reference panel of 64,976 haplotypes for genotype imputation.* Nat Genet, 2016. **48**(10): p. 1279-83.
76. Mitt, M., et al., *Improved imputation accuracy of rare and low-frequency variants using population-specific high-coverage WGS-based imputation reference panel.* Eur J Hum Genet, 2017. **25**(7): p. 869-876.
77. Lohmueller, K.E., et al., *Proportionally more deleterious genetic variation in European than in African populations.* Nature, 2008. **451**(7181): p. 994-7.
78. Casals, F., et al., *Whole-exome sequencing reveals a rapid change in the frequency of rare functional variants in a founding population of humans.* PLoS Genet, 2013. **9**(9): p. e1003815.

79. Lim, E.T., et al., *Distribution and medical impact of loss-of-function variants in the Finnish founder population.* PLoS Genet, 2014. **10**(7): p. e1004494.
80. Gravel, S., *When Is Selection Effective?* Genetics, 2016. **203**(1): p. 451-62.
81. Simons, Y.B., et al., *The deleterious mutation load is insensitive to recent population history.* Nat Genet, 2014. **46**(3): p. 220-4.
82. Do, R., et al., *No evidence that selection has been less effective at removing deleterious mutations in Europeans than in Africans.* Nat Genet, 2015. **47**(2): p. 126-31.
83. Brozek, J.L., et al., *Allergic Rhinitis and its Impact on Asthma (ARIA) guidelines-2016 revision.* J Allergy Clin Immunol, 2017.
84. Subbarao, P., P.J. Mandhane, and M.R. Sears, *Asthma: epidemiology, etiology and risk factors.* CMAJ, 2009. **181**(9): p. E181-90.
85. Bateman, E.D., et al., *Global strategy for asthma management and prevention: GINA executive summary.* Eur Respir J, 2008. **31**(1): p. 143-78.
86. Spergel, J.M., *Epidemiology of atopic dermatitis and atopic march in children.* Immunol Allergy Clin North Am, 2010. **30**(3): p. 269-80.
87. Dharmage, S.C., et al., *Atopic dermatitis and the atopic march revisited.* Allergy, 2014. **69**(1): p. 17-27.
88. Spergel, J.M., *From atopic dermatitis to asthma: the atopic march.* Ann Allergy Asthma Immunol, 2010. **105**(2): p. 99-106; quiz 107-9, 117.
89. Almqvist, C., et al., *Early predictors for developing allergic disease and asthma: examining separate steps in the 'allergic march'.* Clin Exp Allergy, 2007. **37**(9): p. 1296-302.
90. Ricci, G., et al., *Long-term follow-up of atopic dermatitis: retrospective analysis of related risk factors and association with concomitant allergic diseases.* J Am Acad Dermatol, 2006. **55**(5): p. 765-71.
91. Illi, S., et al., *The natural course of atopic dermatitis from birth to age 7 years and the association with asthma.* J Allergy Clin Immunol, 2004. **113**(5): p. 925-31.
92. Pawankar, R., *Allergic rhinitis and asthma: are they manifestations of one syndrome?* Clin Exp Allergy, 2006. **36**(1): p. 1-4.
93. Johansson, S.G. and J. Lundahl, *Asthma, atopy, and IgE: what is the link?* Curr Allergy Asthma Rep, 2001. **1**(2): p. 89-90.
94. Duffy, D.L., et al., *Genetics of asthma and hay fever in Australian twins.* Am Rev Respir Dis, 1990. **142**(6 Pt 1): p. 1351-8.
95. Nieminen, M.M., J. Kaprio, and M. Koskenvuo, *A population-based study of bronchial asthma in adult twin pairs.* Chest, 1991. **100**(1): p. 70-5.
96. Thomsen, S.F., et al., *Estimates of asthma heritability in a large twin sample.* Clin Exp Allergy, 2010. **40**(7): p. 1054-61.
97. Sigurs, N., et al., *Asthma and allergy patterns over 18 years after severe RSV bronchiolitis in the first year of life.* Thorax, 2010. **65**(12): p. 1045-52.
98. Jamieson, K.C., et al., *Rhinovirus in the Pathogenesis and Clinical Course of Asthma.* Chest, 2015. **148**(6): p. 1508-1516.
99. Thomson, N.C. and R. Chaudhuri, *Asthma in smokers: challenges and opportunities.* Curr Opin Pulm Med, 2009. **15**(1): p. 39-45.

100. Maestrelli, P., et al., *Mechanisms of occupational asthma.* J Allergy Clin Immunol, 2009. **123**(3): p. 531-42; quiz 543-4.
101. Thomsen, S.F., et al., *Genetic influence on the age at onset of asthma: a twin study.* J Allergy Clin Immunol, 2010. **126**(3): p. 626-30.
102. van Beijsterveldt, C.E. and D.I. Boomsma, *Genetics of parentally reported asthma, eczema and rhinitis in 5-yr-old twins.* Eur Respir J, 2007. **29**(3): p. 516-21.
103. Fagnani, C., et al., *Heritability and shared genetic effects of asthma and hay fever: an Italian study of young twins.* Twin Res Hum Genet, 2008. **11**(2): p. 121-31.
104. Palmer, L.J., et al., *Independent inheritance of serum immunoglobulin E concentrations and airway responsiveness.* Am J Respir Crit Care Med, 2000. **161**(6): p. 1836-43.
105. Palmer, L.J., et al., *Familial aggregation and heritability of asthma-associated quantitative traits in a population-based sample of nuclear families.* Eur J Hum Genet, 2000. **8**(11): p. 853-60.
106. Barnes, K.C., *An update on the genetics of atopic dermatitis: scratching the surface in 2009.* J Allergy Clin Immunol, 2010. **125**(1): p. 16-29 e1-11; quiz 30-1.
107. Limb, S.L., et al., *Adult asthma severity in individuals with a history of childhood asthma.* J Allergy Clin Immunol, 2005. **115**(1): p. 61-6.
108. Burrows, B., et al., *Association of asthma with serum IgE levels and skin-test reactivity to allergens.* N Engl J Med, 1989. **320**(5): p. 271-7.
109. Hogan, S.P., et al., *Eosinophils: biological properties and role in health and disease.* Clin Exp Allergy, 2008. **38**(5): p. 709-50.
110. Lee, J.J., et al., *Defining a link with asthma in mice congenitally deficient in eosinophils.* Science, 2004. **305**(5691): p. 1773-6.
111. Gudbjartsson, D.F., et al., *Sequence variants affecting eosinophil numbers associate with asthma and myocardial infarction.* Nat Genet, 2009. **41**(3): p. 342-7.
112. Marenholz, I., et al., *Meta-analysis identifies seven susceptibility loci involved in the atopic march.* Nat Commun, 2015. **6**: p. 8804.
113. Almoguera, B., et al., *Identification of Four Novel Loci in Asthma in European American and African American Populations.* Am J Respir Crit Care Med, 2017. **195**(4): p. 456-463.
114. Bonnelykke, K., et al., *Meta-analysis of genome-wide association studies identifies ten loci influencing allergic sensitization.* Nat Genet, 2013. **45**(8): p. 902-6.
115. Ding, L., et al., *Rank-based genome-wide analysis reveals the association of ryanodine receptor-2 gene variants with childhood asthma among human populations.* Hum Genomics, 2013. **7**: p. 16.
116. Ferreira, M.A., et al., *Identification of IL6R and chromosome 11q13.5 as risk loci for asthma.* Lancet, 2011. **378**(9795): p. 1006-14.
117. Forno, E., et al., *Genome-wide association study of the age of onset of childhood asthma.* J Allergy Clin Immunol, 2012. **130**(1): p. 83-90 e4.

118. Himes, B.E., et al., *Genome-wide association analysis identifies PDE4D as an asthma-susceptibility gene.* Am J Hum Genet, 2009. **84**(5): p. 581-93.
119. Hirota, T., et al., *Genome-wide association study identifies three new susceptibility loci for adult asthma in the Japanese population.* Nat Genet, 2011. **43**(9): p. 893-6.
120. Lasky-Su, J., et al., *HLA-DQ strikes again: genome-wide association study further confirms HLA-DQ in the diagnosis of asthma among adults.* Clin Exp Allergy, 2012. **42**(12): p. 1724-33.
121. Moffatt, M.F., et al., *A large-scale, consortium-based genomewide association study of asthma.* N Engl J Med, 2010. **363**(13): p. 1211-21.
122. Moffatt, M.F., et al., *Genetic variants regulating ORMDL3 expression contribute to the risk of childhood asthma.* Nature, 2007. **448**(7152): p. 470-3.
123. Noguchi, E., et al., *Genome-wide association study identifies HLA-DP as a susceptibility gene for pediatric asthma in Asian populations.* PLoS Genet, 2011. **7**(7): p. e1002170.
124. Park, H.W., et al., *Genetic predictors associated with improvement of asthma symptoms in response to inhaled corticosteroids.* J Allergy Clin Immunol, 2014. **133**(3): p. 664-9 e5.
125. Pickrell, J.K., et al., *Detection and interpretation of shared genetic influences on 42 human traits.* Nat Genet, 2016. **48**(7): p. 709-17.
126. Ramasamy, A., et al., *Genome-wide association studies of asthma in population-based cohorts confirm known and suggested loci and identify an additional association near HLA.* PLoS One, 2012. **7**(9): p. e44008.
127. Sleiman, P.M., et al., *Variants of DENND1B associated with asthma in children.* N Engl J Med, 2010. **362**(1): p. 36-44.
128. Torgerson, D.G., et al., *Meta-analysis of genome-wide association studies of asthma in ethnically diverse North American populations.* Nat Genet, 2011. **43**(9): p. 887-92.
129. Wan, Y.I., et al., *Genome-wide association study to identify genetic determinants of severe asthma.* Thorax, 2012. **67**(9): p. 762-8.
130. Yucesoy, B., et al., *Genome-Wide Association Study Identifies Novel Loci Associated With Diisocyanate-Induced Occupational Asthma.* Toxicol Sci, 2015. **146**(1): p. 192-201.
131. Hinds, D.A., et al., *A genome-wide association meta-analysis of self-reported allergy identifies shared and allergy-specific susceptibility loci.* Nat Genet, 2013. **45**(8): p. 907-11.
132. Hong, X., et al., *Genome-wide association study identifies peanut allergy-specific loci and evidence of epigenetic mediation in US children.* Nat Commun, 2015. **6**: p. 6304.
133. Bunyavanich, S., et al., *Integrated genome-wide association, coexpression network, and expression single nucleotide polymorphism analysis identifies novel pathway in allergic rhinitis.* BMC Med Genomics, 2014. **7**: p. 48.
134. Ferreira, M.A., et al., *Genome-wide association analysis identifies 11 risk variants associated with the asthma with hay fever phenotype.* J Allergy Clin Immunol, 2014. **133**(6): p. 1564-71.

135. Ramasamy, A., et al., *A genome-wide meta-analysis of genetic variants associated with allergic rhinitis and grass sensitization and their interaction with birth order.* J Allergy Clin Immunol, 2011. **128**(5): p. 996-1005.
136. Weidinger, S., et al., *A genome-wide association study of atopic dermatitis identifies loci with overlapping effects on asthma and psoriasis.* Hum Mol Genet, 2013. **22**(23): p. 4841-56.
137. Baurecht, H., et al., *Genome-wide comparative analysis of atopic dermatitis and psoriasis gives insight into opposing genetic mechanisms.* Am J Hum Genet, 2015. **96**(1): p. 104-20.
138. Esparza-Gordillo, J., et al., *A common variant on chromosome 11q13 is associated with atopic dermatitis.* Nat Genet, 2009. **41**(5): p. 596-601.
139. Hirota, T., et al., *Genome-wide association study identifies eight new susceptibility loci for atopic dermatitis in the Japanese population.* Nat Genet, 2012. **44**(11): p. 1222-6.
140. Kim, K.W., et al., *Genome-wide association study of recalcitrant atopic dermatitis in Korean children.* J Allergy Clin Immunol, 2015. **136**(3): p. 678-684 e4.
141. Paternoster, L., et al., *Meta-analysis of genome-wide association studies identifies three new risk loci for atopic dermatitis.* Nat Genet, 2011. **44**(2): p. 187-92.
142. Paternoster, L., et al., *Multi-ancestry genome-wide association study of 21,000 cases and 95,000 controls identifies new risk loci for atopic dermatitis.* Nat Genet, 2015. **47**(12): p. 1449-1456.
143. Schaarschmidt, H., et al., *A genome-wide association study reveals 2 new susceptibility loci for atopic dermatitis.* J Allergy Clin Immunol, 2015. **136**(3): p. 802-6.
144. Sun, L.D., et al., *Genome-wide association study identifies two new susceptibility loci for atopic dermatitis in the Chinese Han population.* Nat Genet, 2011. **43**(7): p. 690-4.
145. Granada, M., et al., *A genome-wide association study of plasma total IgE concentrations in the Framingham Heart Study.* J Allergy Clin Immunol, 2012. **129**(3): p. 840-845 e21.
146. Weidinger, S., et al., *Genome-wide scan on total serum IgE levels identifies FCER1A as novel susceptibility locus.* PLoS Genet, 2008. **4**(8): p. e1000166.
147. Yatagai, Y., et al., *Genome-wide association study for levels of total serum IgE identifies HLA-C in a Japanese population.* PLoS One, 2013. **8**(12): p. e80941.
148. Jain, D., et al., *Genome-wide association of white blood cell counts in Hispanic/Latino Americans: the Hispanic Community Health Study/Study of Latinos.* Hum Mol Genet, 2017. **26**(6): p. 1193-1204.
149. Okada, Y., et al., *Identification of nine novel loci associated with white blood cell subtypes in a Japanese population.* PLoS Genet, 2011. **7**(6): p. e1002067.
150. Forno, E., et al., *Genome-wide interaction study of dust mite allergen on lung function in children with asthma.* J Allergy Clin Immunol, 2017.
151. Hancock, D.B., et al., *Meta-analyses of genome-wide association studies identify multiple loci associated with pulmonary function.* Nat Genet, 2010. **42**(1): p. 45-52.

152. Ober, C., et al., *Effect of variation in CHI3L1 on serum YKL-40 level, risk of asthma, and lung function.* N Engl J Med, 2008. **358**(16): p. 1682-91.
153. Repapi, E., et al., *Genome-wide association study identifies five loci associated with lung function.* Nat Genet, 2010. **42**(1): p. 36-44.
154. Soler Artigas, M., et al., *Genome-wide association and large-scale follow up identifies 16 new loci influencing lung function.* Nat Genet, 2011. **43**(11): p. 1082-90.
155. Wilk, J.B., et al., *A genome-wide association study of pulmonary function measures in the Framingham Heart Study.* PLoS Genet, 2009. **5**(3): p. e1000429.
156. Wilk, J.B., et al., *Genome-wide association studies identify CHRNA5/3 and HTR4 in the development of airflow obstruction.* Am J Respir Crit Care Med, 2012. **186**(7): p. 622-32.
157. Yao, T.C., et al., *Genome-wide association study of lung function phenotypes in a founder population.* J Allergy Clin Immunol, 2014. **133**(1): p. 248-55 e1-10.
158. Torgerson, D.G., et al., *Resequencing candidate genes implicates rare variants in asthma susceptibility.* Am J Hum Genet, 2012. **90**(2): p. 273-81.
159. Drake, K.A., et al., *A genome-wide association study of bronchodilator response in Latinos implicates rare variants.* J Allergy Clin Immunol, 2014. **133**(2): p. 370-8.
160. Torgerson, D.G., et al., *Pooled Sequencing of Candidate Genes Implicates Rare Variants in the Development of Asthma Following Severe RSV Bronchiolitis in Infancy.* PLoS One, 2015. **10**(11): p. e0142649.
161. Wain, L.V., et al., *Novel insights into the genetics of smoking behaviour, lung function, and chronic obstructive pulmonary disease (UK BiLEVE): a genetic association study in UK Biobank.* Lancet Respir Med, 2015. **3**(10): p. 769-81.
162. Ellinghaus, D., et al., *High-density genotyping study identifies four new susceptibility loci for atopic dermatitis.* Nat Genet, 2013. **45**(7): p. 808-12.
163. Gao, L., et al., *Targeted deep sequencing identifies rare loss-of-function variants in IFNGR1 for risk of atopic dermatitis complicated by eczema herpeticum.* J Allergy Clin Immunol, 2015. **136**(6): p. 1591-600.
164. Hao, K., et al., *Lung eQTLs to help reveal the molecular underpinnings of asthma.* PLoS Genet, 2012. **8**(11): p. e1003029.
165. Luo, W., et al., *Airway Epithelial Expression Quantitative Trait Loci Reveal Genes Underlying Asthma and Other Airway Diseases.* Am J Respir Cell Mol Biol, 2016. **54**(2): p. 177-87.
166. Sharma, S., et al., *A genome-wide survey of CD4(+) lymphocyte regulatory genetic variants identifies novel asthma genes.* J Allergy Clin Immunol, 2014. **134**(5): p. 1153-62.
167. Li, X., et al., *eQTL of bronchial epithelial cells and bronchial alveolar lavage deciphers GWAS-identified asthma genes.* Allergy, 2015. **70**(10): p. 1309-18.
168. 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.

169. Lluis, A., et al., *Asthma-associated polymorphisms in 17q21 influence cord blood ORMDL3 and GSDMA gene expression and IL-17 secretion.* J Allergy Clin Immunol, 2011. **127**(6): p. 1587-94 e6.
170. Halapi, E., et al., *A sequence variant on 17q21 is associated with age at onset and severity of asthma.* Eur J Hum Genet, 2010. **18**(8): p. 902-8.
171. Gordon, E.D., et al., *IL1RL1 asthma risk variants regulate airway type 2 inflammation.* JCI Insight, 2016. **1**(14): p. e87871.
172. Akhabir, L., et al., *Lung expression quantitative trait loci data set identifies important functional polymorphisms in the asthma-associated IL1RL1 region.* J Allergy Clin Immunol, 2014. **134**(3): p. 729-31.
173. Berube, J.C., et al., *Identification of Susceptibility Genes of Adult Asthma in French Canadian Women.* Can Respir J, 2016. **2016**: p. 3564341.
174. Ferreira, M.A., et al., *Gene-based analysis of regulatory variants identifies 4 putative novel asthma risk genes related to nucleotide synthesis and signaling.* J Allergy Clin Immunol, 2017. **139**(4): p. 1148-1157.
175. Nieuwenhuis, M.A., et al., *Combining genomewide association study and lung eQTL analysis provides evidence for novel genes associated with asthma.* Allergy, 2016. **71**(12): p. 1712-1720.
176. Himes, B.E., et al., *ITGB5 and AGFG1 variants are associated with severity of airway responsiveness.* BMC Med Genet, 2013. **14**: p. 86.
177. Babron, M.C., et al., *Rare and low frequency variant stratification in the UK population: description and impact on association tests.* PLoS One, 2012. **7**(10): p. e46519.
178. Price, A.L., et al., *Principal components analysis corrects for stratification in genome-wide association studies.* Nat Genet, 2006. **38**(8): p. 904-9.
179. Mathieson, I. and G. McVean, *Differential confounding of rare and common variants in spatially structured populations.* Nat Genet, 2012. **44**(3): p. 243-6.
180. Hatzikotoulas, K., A. Gilly, and E. Zeggini, *Using population isolates in genetic association studies.* Brief Funct Genomics, 2014. **13**(5): p. 371-7.
181. Rudan, I., *Health effects of human population isolation and admixture.* Croat Med J, 2006. **47**(4): p. 526-31.
182. De Braekeleer, M., *Geographic distribution of 18 autosomal recessive disorders in the French Canadian population of Saguenay-Lac-Saint-Jean, Quebec.* Ann Hum Biol, 1995. **22**(2): p. 111-22.
183. Moltke, I., et al., *A common Greenlandic TBC1D4 variant confers muscle insulin resistance and type 2 diabetes.* Nature, 2014. **512**(7513): p. 190-3.
184. Zoledziewska, M., et al., *Height-reducing variants and selection for short stature in Sardinia.* Nat Genet, 2015. **47**(11): p. 1352-6.
185. Igartua, C., et al., *Rare non-coding variants are associated with plasma lipid traits in a founder population.* bioRxiv, 2017.
186. Laprise, C., *The Saguenay-Lac-Saint-Jean asthma familial collection: the genetics of asthma in a young founder population.* Genes Immun, 2014. **15**(4): p. 247-55.
187. Madore, A.M., et al., *Genes Involved in Interleukin-1 Receptor Type II Activities Are Associated With Asthmatic Phenotypes.* Allergy Asthma Immunol Res, 2016. **8**(5): p. 466-70.

188. Bosse, Y., et al., *Asthma and genes encoding components of the vitamin D pathway.* Respir Res, 2009. **10**: p. 98.
189. Tremblay, K., et al., *Association study between the CX3CR1 gene and asthma.* Genes Immun, 2006. **7**(8): p. 632-9.
190. Begin, P., et al., *Association of urokinase-type plasminogen activator with asthma and atopy.* Am J Respir Crit Care Med, 2007. **175**(11): p. 1109-16.
191. Poon, A., et al., *ATG5, autophagy and lung function in asthma.* Autophagy, 2012. **8**(4): p. 694-5.
192. Gagne-Ouellet, V., et al., *DNA methylation signature of interleukin 1 receptor type II in asthma.* Clin Epigenetics, 2015. **7**: p. 80.
193. Moussette, S., et al., *Role of DNA methylation in expression control of the IKZF3-GSDMA region in human epithelial cells.* PLoS One, 2017. **12**(2): p. e0172707.
194. Al Tuwaijri, A., et al., *Local genotype influences DNA methylation at two asthma-associated regions, 5q31 and 17q21, in a founder effect population.* J Med Genet, 2016. **53**(4): p. 232-41.
195. Verlaan, D.J., et al., *Allele-specific chromatin remodeling in the ZPBP2/GSDMB/ORMDL3 locus associated with the risk of asthma and autoimmune disease.* Am J Hum Genet, 2009. **85**(3): p. 377-93.
196. Naumova, A.K., et al., *Sex- and age-dependent DNA methylation at the 17q12-q21 locus associated with childhood asthma.* Hum Genet, 2013. **132**(7): p. 811-22.
197. Liang, L., et al., *An epigenome-wide association study of total serum immunoglobulin E concentration.* Nature, 2015. **520**(7549): p. 670-4.
198. Visscher, P.M., et al., *Five years of GWAS discovery.* Am J Hum Genet, 2012. **90**(1): p. 7-24.
199. Zheng, H.F., et al., *Performance of genotype imputation for low frequency and rare variants from the 1000 genomes.* PLoS One, 2015. **10**(1): p. e0116487.
200. McCarthy, M.I., et al., *Genome-wide association studies for complex traits: consensus, uncertainty and challenges.* Nat Rev Genet, 2008. **9**(5): p. 356-69.
201. Do, R., et al., *Exome sequencing identifies rare LDLR and APOA5 alleles conferring risk for myocardial infarction.* Nature, 2015. **518**(7537): p. 102-6.
202. Seddon, J.M., et al., *Rare variants in CFI, C3 and C9 are associated with high risk of advanced age-related macular degeneration.* Nat Genet, 2013. **45**(11): p. 1366-70.
203. Helgason, H., et al., *A rare nonsynonymous sequence variant in C3 is associated with high risk of age-related macular degeneration.* Nat Genet, 2013. **45**(11): p. 1371-4.
204. Farh, K.K., et al., *Genetic and epigenetic fine mapping of causal autoimmune disease variants.* Nature, 2015. **518**(7539): p. 337-43.
205. Ernst, J., et al., *Mapping and analysis of chromatin state dynamics in nine human cell types.* Nature, 2011. **473**(7345): p. 43-9.
206. Consortium, E.P., et al., *An integrated encyclopedia of DNA elements in the human genome.* Nature, 2012. **489**(7414): p. 57-74.
207. Roadmap Epigenomics, C., et al., *Integrative analysis of 111 reference human epigenomes.* Nature, 2015. **518**(7539): p. 317-30.

208. Lappalainen, T., et al., *Transcriptome and genome sequencing uncovers functional variation in humans.* Nature, 2013. **501**(7468): p. 506-11.
209. Gaffney, D.J., et al., *Dissecting the regulatory architecture of gene expression QTLs.* Genome Biol, 2012. **13**(1): p. R7.
210. Fairfax, B.P., et al., *Genetics of gene expression in primary immune cells identifies cell type-specific master regulators and roles of HLA alleles.* Nat Genet, 2012. **44**(5): p. 502-10.
211. Adoue, V., et al., *Allelic expression mapping across cellular lineages to establish impact of non-coding SNPs.* Mol Syst Biol, 2014. **10**: p. 754.
212. Battle, A., et al., *Characterizing the genetic basis of transcriptome diversity through RNA-sequencing of 922 individuals.* Genome Res, 2014. **24**(1): p. 14-24.
213. Montgomery, S.B., et al., *Rare and common regulatory variation in population-scale sequenced human genomes.* PLoS Genet, 2011. **7**(7): p. e1002144.
214. Chadwick, L.H., *The NIH Roadmap Epigenomics Program data resource.* Epigenomics, 2012. **4**(3): p. 317-24.
215. 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 Res, 2012. **40**(Database issue): p. D930-4.
216. 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.
217. Drmanac, R., et al., *Human genome sequencing using unchained base reads on self-assembling DNA nanoarrays.* Science, 2010. **327**(5961): p. 78-81.
218. Davydov, E.V., et al., *Identifying a high fraction of the human genome to be under selective constraint using GERP++.* PLoS Comput Biol, 2010. **6**(12): p. e1001025.
219. Kircher, M., et al., *A general framework for estimating the relative pathogenicity of human genetic variants.* Nat Genet, 2014. **46**(3): p. 310-5.
220. Cingolani, P., et al., *A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3.* Fly (Austin), 2012. **6**(2): p. 80-92.
221. Bolger, A.M., M. Lohse, and B. Usadel, *Trimmomatic: a flexible trimmer for Illumina sequence data.* Bioinformatics, 2014. **30**(15): p. 2114-20.
222. Trapnell, C., L. Pachter, and S.L. Salzberg, *TopHat: discovering splice junctions with RNA-Seq.* Bioinformatics, 2009. **25**(9): p. 1105-11.
223. Langmead, B., et al., *Ultrafast and memory-efficient alignment of short DNA sequences to the human genome.* Genome Biol, 2009. **10**(3): p. R25.
224. Lemire, M., et al., *Long-range epigenetic regulation is conferred by genetic variation located at thousands of independent loci.* Nat Commun, 2015. **6**: p. 6326.
225. Rao, S.S., et al., *A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping.* Cell, 2014. **159**(7): p. 1665-80.
226. Natarajan, A., et al., *Predicting cell-type-specific gene expression from regions of open chromatin.* Genome Res, 2012. **22**(9): p. 1711-22.

227. Panoutsopoulou, K., I. Tachmazidou, and E. Zeggini, *In search of low-frequency and rare variants affecting complex traits.* Hum Mol Genet, 2013. **22**(R1): p. R16-21.
228. Xue, Y., et al., *Enrichment of low-frequency functional variants revealed by whole-genome sequencing of multiple isolated European populations.* Nat Commun, 2017. **8**: p. 15927.
229. Morin, A., et al., *Immunoseq: the identification of functionally relevant variants through targeted capture and sequencing of active regulatory regions in human immune cells.* BMC Med Genomics, 2016. **9**(1): p. 59.
230. 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.
231. Zheng, X., et al., *A high-performance computing toolset for relatedness and principal component analysis of SNP data.* Bioinformatics, 2012. **28**(24): p. 3326-8.
232. Danecek, P., et al., *The variant call format and VCFtools.* Bioinformatics, 2011. **27**(15): p. 2156-8.
233. Delaneau, O., et al., *Integrating sequence and array data to create an improved 1000 Genomes Project haplotype reference panel.* Nat Commun, 2014. **5**: p. 3934.
234. Delaneau, O., J. Marchini, and J.F. Zagury, *A linear complexity phasing method for thousands of genomes.* Nat Methods, 2012. **9**(2): p. 179-81.
235. Delaneau, O., J.F. Zagury, and J. Marchini, *Improved whole-chromosome phasing for disease and population genetic studies.* Nat Methods, 2013. **10**(1): p. 5-6.
236. Hussin, J., et al., *Age-dependent recombination rates in human pedigrees.* PLoS Genet, 2011. **7**(9): p. e1002251.
237. Howie, B.N., P. Donnelly, and J. Marchini, *A flexible and accurate genotype imputation method for the next generation of genome-wide association studies.* PLoS Genet, 2009. **5**(6): p. e1000529.
238. Li, B. and S.M. Leal, *Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data.* Am J Hum Genet, 2008. **83**(3): p. 311-21.
239. Ionita-Laza, I., et al., *Sequence kernel association tests for the combined effect of rare and common variants.* Am J Hum Genet, 2013. **92**(6): p. 841-53.
240. Ferland, C., et al., *Eotaxin promotes eosinophil transmigration via the activation of the plasminogen-plasmin system.* J Leukoc Biol, 2001. **69**(5): p. 772-8.
241. Morin, A., et al., *Combining omics data to identify genes associated with allergic rhinitis.* Clin Epigenetics, 2017. **9**: p. 3.
242. Pankiv, S. and T. Johansen, *FYCO1: linking autophagosomes to microtubule plus end-directing molecular motors.* Autophagy, 2010. **6**(4): p. 550-2.
243. Jyothula, S.S. and N.T. Eissa, *Autophagy and role in asthma.* Curr Opin Pulm Med, 2013. **19**(1): p. 30-5.
244. Renz, H., *Autophagy: Nobel Prize 2016 and allergy and asthma research.* J Allergy Clin Immunol, 2017.

245. Poon, A.H., et al., *Increased Autophagy-Related 5 Gene Expression Is Associated with Collagen Expression in the Airways of Refractory Asthmatics.* Front Immunol, 2017. **8**: p. 355.
246. Martin, L.J., et al., *Functional variant in the autophagy-related 5 gene promotor is associated with childhood asthma.* PLoS One, 2012. **7**(4): p. e33454.
247. Latta, M., K. Mohan, and T.B. Issekutz, *CXCR6 is expressed on T cells in both T helper type 1 (Th1) inflammation and allergen-induced Th2 lung inflammation but is only a weak mediator of chemotaxis.* Immunology, 2007. **121**(4): p. 555-64.
248. Tian, D.B., et al., *Association between the rs3795879 G/A polymorphism of the SERPINE2 gene and chronic obstructive pulmonary disease: a meta-analysis.* Genet Mol Res, 2015. **14**(3): p. 7920-8.
249. Fujimoto, K., et al., *Polymorphism of SERPINE2 gene is associated with pulmonary emphysema in consecutive autopsy cases.* BMC Med Genet, 2010. **11**: p. 159.
250. Himes, B.E., et al., *Association of SERPINE2 with asthma.* Chest, 2011. **140**(3): p. 667-674.
251. Jung, J., et al., *Multiplex image-based autophagy RNAi screening identifies SMCR8 as ULK1 kinase activity and gene expression regulator.* Elife, 2017. **6**.
252. Bousquet, J., et al., *Allergic Rhinitis and its Impact on Asthma (ARIA) 2008 update (in collaboration with the World Health Organization, GA(2)LEN and AllerGen).* Allergy, 2008. **63 Suppl 86**: p. 8-160.
253. Berger, S.L., *The complex language of chromatin regulation during transcription.* Nature, 2007. **447**(7143): p. 407-12.
254. Hannon, E., et al., *Methylation QTLs in the developing brain and their enrichment in schizophrenia risk loci.* Nat Neurosci, 2016. **19**(1): p. 48-54.
255. Banovich, N.E., et al., *Methylation QTLs are associated with coordinated changes in transcription factor binding, histone modifications, and gene expression levels.* PLoS Genet, 2014. **10**(9): p. e1004663.
256. Bell, J.T., et al., *Epigenome-wide scans identify differentially methylated regions for age and age-related phenotypes in a healthy ageing population.* PLoS Genet, 2012. **8**(4): p. e1002629.
257. McClay, J.L., et al., *High density methylation QTL analysis in human blood via next-generation sequencing of the methylated genomic DNA fraction.* Genome Biol, 2015. **16**: p. 291.
258. Madore, A.M., et al., *Contribution of hierarchical clustering techniques to the modeling of the geographic distribution of genetic polymorphisms associated with chronic inflammatory diseases in the Quebec population.* Community Genet, 2007. **10**(4): p. 218-26.
259. Madore, A.M., et al., *Distribution of CFTR mutations in Saguenay- Lac-Saint-Jean: proposal of a panel of mutations for population screening.* Genet Med, 2008. **10**(3): p. 201-6.
260. Thornton, T. and M.S. McPeek, *Case-control association testing with related individuals: a more powerful quasi-likelihood score test.* Am J Hum Genet, 2007. **81**(2): p. 321-37.

261. Assenov, Y., et al., *Comprehensive analysis of DNA methylation data with RnBeads.* Nat Methods, 2014. **11**(11): p. 1138-40.
262. Davis S, D.P., Bilke S, Triche T, Jr. and Bootwalla M *methylumi: Handle Illumina methylation data*, 2015.
263. Aryee, M.J., et al., *Minfi: a flexible and comprehensive Bioconductor package for the analysis of Infinium DNA methylation microarrays.* Bioinformatics, 2014. **30**(10): p. 1363-9.
264. Maksimovic, J., L. Gordon, and A. Oshlack, *SWAN: Subset-quantile within array normalization for illumina infinium HumanMethylation450 BeadChips.* Genome Biol, 2012. **13**(6): p. R44.
265. Smyth, G.K., *Limma: linear models for microarray data*, in *Bioinformatics and Computational Biology Solutions using R and Bioconductor*, V.C. R. Gentleman, S. Dudoit, R. Irizarry, W. Huber,, Editor. 2005, Springer: New York.
266. Shabalin, A.A., *Matrix eQTL: ultra fast eQTL analysis via large matrix operations.* Bioinformatics, 2012. **28**(10): p. 1353-8.
267. Hancock, D.B., et al., *Genome-wide joint meta-analysis of SNP and SNP-by-smoking interaction identifies novel loci for pulmonary function.* PLoS Genet, 2012. **8**(12): p. e1003098.
268. Li, X., et al., *Genome-wide association studies of asthma indicate opposite immunopathogenesis direction from autoimmune diseases.* J Allergy Clin Immunol, 2012. **130**(4): p. 861-8 e7.
269. Lukacs, N.W., *Role of chemokines in the pathogenesis of asthma.* Nat Rev Immunol, 2001. **1**(2): p. 108-16.
270. Laprise, C., et al., *Functional classes of bronchial mucosa genes that are differentially expressed in asthma.* BMC Genomics, 2004. **5**(1): p. 21.
271. Madore, A.M., et al., *Alveolar macrophages in allergic asthma: an expression signature characterized by heat shock protein pathways.* Hum Immunol, 2010. **71**(2): p. 144-50.
272. Wenzel, S.E., *Asthma phenotypes: the evolution from clinical to molecular approaches.* Nat Med, 2012. **18**(5): p. 716-25.
273. Bisgaard, H., et al., *Chromosome 17q21 gene variants are associated with asthma and exacerbations but not atopy in early childhood.* Am J Respir Crit Care Med, 2009. **179**(3): p. 179-85.
274. Poon, A.H., et al., *Pathogenesis of severe asthma.* Clin Exp Allergy, 2012. **42**(5): p. 625-37.
275. Bonnelykke, K., et al., *A genome-wide association study identifies CDHR3 as a susceptibility locus for early childhood asthma with severe exacerbations.* Nat Genet, 2014. **46**(1): p. 51-5.
276. Moore, W.C., et al., *Identification of asthma phenotypes using cluster analysis in the Severe Asthma Research Program.* Am J Respir Crit Care Med, 2010. **181**(4): p. 315-23.
277. Caliskan, M., et al., *Rhinovirus wheezing illness and genetic risk of childhood-onset asthma.* N Engl J Med, 2013. **368**(15): p. 1398-407.
278. Loss, G.J., et al., *The Early Development of Wheeze. Environmental Determinants and Genetic Susceptibility at 17q21.* Am J Respir Crit Care Med, 2016. **193**(8): p. 889-97.

279. Bonnelykke, K. and C. Ober, *Leveraging gene-environment interactions and endotypes for asthma gene discovery.* J Allergy Clin Immunol, 2016. **137**(3): p. 667-79.

Appendix

Significant Contribution by the author to other projects

Published work

1. Sugier PE, Brossard M, Sarnowski C, Vaysse A, **Morin A**, Pain L, Margaritte-Jeannin P, Dizier MH, Cookson WOCM, Lathrop M, Moffatt MF, Laprise C, Demenais F, Bouzigon E. A novel role for ciliary function in atopy: ADGRV1 and DNAH5 interactions. J Allergy Clin Immunol. 2017 Sept 18; PMID:28927820

2. Cheung WA*, Shao XJ*, **Morin A***, Siroux V, Kwan T, Ge B, Aïssi D, Chen L, Vasquez L, Allum F, Guénard F, Bouzigon E, Simon MM, Boulier EL, Redensek A, Watt S, Datta A, Clarke L, Flicek P, Mead D, Paul DS, Beck S, Bourque G, Lathrop M, Tchernof A, Vohl MC, Demenais F, Pin I, Downes K, Stunnenberg HG, Soranzo N, Grundberg E, Pastinen T, *Functional variation in allelic methylomes underscore strong genetic contribution and reveal novel epigenetic alterations in the human epigenome.* Genome Biol. 2017 Mar 10;18(1):50. PMID: 28283040, * co-first authors

3. Hocking TD, Goerner-Potvin P, **Morin A**, Shao X, Pastinen T, Bourque G. *Optimizing ChIP-seq peak detectors using visual labels and supervised machine learning*. Bioinformatics. 2016 Oct 24. PMID: 27797775

4. Sarnowski C, Laprise C, Malerba G, Moffatt MF, Dizier MH, **Morin A**, Vincent QB, Rohde K, Esparza-Gordillo J, Margaritte-Jeannin P, Liang L, Lee YA, Bousquet J, Siroux V, Pignatti PF, Cookson WO, Lathrop M, Pastinen T, Demenais F, Bouzigon E. *DNA methylation within melatonin receptor 1A (MTNR1A) mediates paternally transmitted genetic variant effect on asthma plus rhinitis.* J Allergy Clin Immunol. 2016 Sep;138(3):748-53. PMID: 27038909

5. Siuko M, Valori M, Kivelä T, Setälä K, **Morin A**, Kwan T, Pastinen T, Tienari P. *Exome and regulatory element sequencing of neuromyelitis optica patients*. J Neuroimmunol. 2015 Dec 15;289:139-42. PMID: 26616883

Manuscript submitted or in preparation

1. Xu CJ, Soderhall C, Bustamante M, Baiz N, Gruzieva O, Gehring U, Mason D, Chatzi L, Basteerechea M, Llop S, Torrent M, Forastiere F, Fantini MP, Lodrup Carlsen KC, Haahtela T, **Morin A**, Kerkhof M, Merid SK, van Rijkom B, Jankipersadsing SA, Bonder MJ, Ballereau S, Vermeulen CJ, Aguirre-Gamboa R, de Jongste JC, Smit HA, Kumar A, Pershagen G, Guerra S, Garcia-Aymerich J, Greco D, Reinius L, McEachan RRC, Azad R, Hovlan V, Mowinckel P, Alenius H, Fyhrquist N, Lemonnier N, Pellet J, Auffray C, the BIOS Consortium, van der Vlies P, van Diemen CC, Li Y, Wijmenga C, Netea MG, Moffatt MF, Cookson WOCM, Anto JM, Bousquet J, Laatikainen T, Laprise C, Carlsen KH, Gori D, Porta D, I.iguez C, Ramon Bilbao J, Kogevinas M, Wright J, Brunekreef B, Kere J, Nawijn MC,

Annesi-Maesano I, Sunyer J, Melen E and Koppelman GH, *Epigenome wide meta-analysis identifies reduced DNA methylation reflecting eosinophil and T cell gene expression signature in childhood asthma*. Accepted in Lancet Respiratory Medicine, Dec 2017

2. Stein MM, Thompson EE, Schoettler N, Helling BA, Magnaye KM, Stanhope C, Igartua C, **Morin A**, Washington III C, Nicolae D, Bønnelykke K, Ober C, A Decade of Research on the 17q12-21 Asthma Locus: Piecing Together the Puzzle. Accepted in J Allergy Clin Immunol.

3. Pain L*, **Morin A***, Boucher-Lafleur AM*, Meloche J and Laprise C, *The Saguenay–Lac-Saint-Jean asthma familial collection: contributions to the omic landscape of asthma,* * co-first authors

4. **Morin A**, Pastinen T, *Exploring rare coding and non-coding variants reveals new genes associated with autoimmune diseases.*