UNIVERSITY OF CALIFORNIA,
IRVINE

Genetics of Celiac Disease

DISSERTATION

submitted in partial satisfaction of the requirements
for the degree of

DOCTOR OF PHILOSOPHY

in Epidemiology

by

Richard Sungho Ahn

Dissertation Committee:
Professor Chad Garner, Chair
Professor Daniel Gillen
Professor Xiaohui Xie

2014

DEDICATION

To

the future pioneers of science.

May our present work help you see just that much further.

TABLE OF CONTENTS

	Page
LIST OF FIGURES	iv
LIST OF TABLES	v
ACKNOWLEDGMENTS	vii
CURRICULUM VITAE	viii
ABSTRACT OF THE DISSERTATION	x
CHAPTER 1: Introduction	1
CHAPTER 2: Methods and Materials	11
CHAPTER 3: Meta-analysis Methods for Genome-wide Association Studies	20
CHAPTER 4: Association Analysis of the Extended MHC Region in Celiac Disease	42
CHAPTER 5: Identification of Rare and Low-frequency Variants Associated with Celiac Disease in the 12 Previously Identified Regions and the MHC Region	62
CHAPTER 6: Analysis of Imputed Low-frequency and Rare Variants	86
CHAPTER 7: Discussion and Conclusions	111
REFERENCES	119

LIST OF FIGURES

		Page
Figure 2.1	Criteria for diagnosing celiac disease	16
Figure 2.2	Stages of previous celiac disease GWAS	19
Figure 3.1	P-M plot of rs13151961	31
Figure 3.2	P-M plot of rs917997	32
Figure 3.3	P-M plot of rs1464510	32
Figure 3.4	P-M plot of rs10903122	33
Figure 3.5	P-M plot of rs1050976	36
Figure 3.6	P-M plot of rs3184504	36
Figure 3.7	P-M plot of 2030519	37
Figure 4.1	Unadjusted association results across the xMHC	49
Figure 4.2	Binary tree computed by conditional recursive partitioning	51
Figure 4.3A	Adjusted association results across full xMHC	53
Figure 4.3B	Adjusted association results focused on HLA genes	53
Figure 4.4	Results of sensitivity analysis	55
Figure 4.5	Adjusted association analysis after SNP grouping	58
Figure 6.1	Flowchart to impute low-frequency and rare variants	88
Figure 6.2	Power to detect an association as a function of the minor allele frequency	106

LIST OF TABLES

		Page
Table 2.1	Number of cases and controls by institution	12
Table 2.2	Number of cases and controls by country for stage 1	15
Table 2.3	Number of cases and controls by country for stage 2	15
Table 2.4	Number of cases and controls by country from Trynka et al. 2011	15
Table 3.1	Meta-analysis p-values from Dubois et al. 2010	29
Table 3.2	Meta-analysis p-values from Trynka et al. 2011	34
Table 3.3	Minor alleles, allele frequencies, and odds ratios for three representative SNPs	38
Table 4.1	HLA high-risk genotypes factorized into a five-level variable	52
Table 4.2	Seven index SNPs	57
Table 5.1	Twelve non-MHC regions that were resequenced	75
Table 5.2	Count of variants by type	78
Table 5.3	Count of novel variants	78
Table 5.4	Results of C-alpha test of MHC genes (MAF $\leq$ 0.04)	80
Table 5.5	Results of C-alpha test of NS variants in MHC genes (MAF $\leq$ 0.04)	81
Table 6.1	Number of successfully imputed variants	99
Table 6.2	Proportion of well-imputed low-frequency or rare variants	99
Table 6.3	Results of burden test (MAF $<$ 0.01)	101
Table 6.4	Results of burden test (MAF $\leq$ 0.04)	101
Table 6.5	Results of C-alpha test (MAF $<$ 0.01)	101
Table 6.6	Results of C-alpha test (MAF $\leq$ 0.04)	102
Table 6.7	Results of C-alpha test including only NS variants (MAF $<$ 0.01)	103
Table 6.8	Results of C-alpha test including only NS variants (MAF $\leq$ 0.04)	104

Table 6.9 Single-marker association results for low-frequency and rare variants 105

ACKNOWLEDGMENTS

I would like to express my deepest appreciation to my committee chair, Professor Chad Garner. Dr. Garner has believed in me from day one and has provided me with the utmost support and guidance as a mentor and supervisor over the past four years. He has constantly challenged me to push myself beyond my current limits and as a result, I am a better scientist and scholar today.

I would like to thank my committee members, Professors Daniel Gillen and Xiaohui Xie, who have provided invaluable instruction, advice, and feedback throughout my tenure as a graduate student. I also want to thank Professors Al Ziogas and Ralph Delfino for their feedback and insight during my advancement to candidacy.

I would also like to thank Professor Susan Neuhausen and her lab at the City of Hope for providing me with experimental data and support in the preparation of my manuscript. In addition, I would like to thank Professor David van Heel at Queen Mary, University of London for generously sharing his lab's genotype data with me.

I want to offer a special thank you to Dr. Bill Karnes of the UC Irvine Medical Center for allowing me to shadow him during his rounds and learn more about the clinical aspects of celiac disease and to Professor Dana Mukamel of the Health Policy Research Institute for her mentorship and support from the time of my employment in her group until the present time.

I would like to thank all the members of the Department of Epidemiology for providing a supportive environment in which to work and grow over the past four years. In particular, I thank Julie Strope, Julia Hernandez, Martha DeYoung, Professor Thomas Taylor, and Professor Hoda Anton-Culver for all of their support and guidance. I also thank my fellow graduate students from the department for enlivening discussions at our weekly journal club and seminar meetings.

I would like to thank Professors John Billimek and Gregory Weiss for the weekly bike rides, life mentoring, and friendship. I would like to thank Professor Josh Swamidass of Washington University for encouraging me to pursue a PhD in the first place. I also thank my friends Sharine Wittkopp, Mona Wood, and Sepehr Akhavan for being great classmates, writing partners, and for providing a safe place to air out new ideas and vent frustrations.

Finally, I would like to thank my parents, Yong & Sunsook Ahn, mother-in-law, Joan McKittrick, father-in-law, Ed Bonlaron, step-father-in-law, Bill Milman, grandparents, David & Louise McKittrick, sister, Jin Ahn, sister-in-law, Lisa Bonlaron, and most importantly, my wonderful wife, Joy Ahn for their unfailing love and support throughout the past four years. I couldn't have finished this manuscript without them, especially without the constant, unconditional support from my wife—I cannot thank her enough!

Financial support was provided by the University of California, Irvine, NIH grants R01 DK081645 and the Department of Epidemiology. The content of this manuscript is solely the responsibility of the author and does not represent the official views of the NIH.

CURRICULUM VITAE

Richard Sungho Ahn

EDUCATION

University of California, Irvine, School of Medicine, Irvine, CA 2010-2014
Doctor of Philosophy, Epidemiology

San Diego State University, San Diego, CA 2005-2006
Master of Arts, Economics

University of California, San Diego, La Jolla, CA 2000-2005
Bachelor of Arts, History and Economics (minor)

RESEARCH EXPERIENCE

University of California, Irvine, Department of Epidemiology, Irvine, CA 2010-2013
Graduate Student Researcher
Advisor: Chad Garner, DPhil
Conducted GWAS, LD-based fine-mapping, developed a pipeline for calling and annotating variants from next-generation sequencing data, rare variant association testing, and contributed to the preparation of manuscripts for the GWAS of Celiac Disease project.

University of California, Irvine, Health Policy Research Institute, Irvine, CA 2008-2010
Statistician
Supervisor: Dana Mukamel, Ph.D.
Conducted primary statistical analysis of large-scale institutional and patient-level datasets from skilled nursing facilities and contributed to the preparation of manuscripts.

PUBLICATIONS

Refereed Journal Articles

1. **Ahn R.**, Garner C. Meta-analysis of Genome-wide Association Studies (GWAS) in Celiac Disease: A Case Study Comparing Fixed and Random Effects Models. *In preparation.*
2. Garner C., **Ahn R.**, Ding Y. C., Steele L., Green P., Fasano A., Murray J., Neuhausen S. L. Genome-wide Association Study of Celiac Disease in North America Confirms FRMD4B as New Celiac Locus. PLOS ONE, 2014. *In review.*
3. **Ahn R.**, Ding Y. C., Murray J., Fasano A., Green P., Neuhausen S. L., Garner C. Association Analysis of the Extended MHC Region in Celiac Disease Implicates Multiple Independent Susceptibility. PLOS ONE 7(5):e36926, 2012.
4. Mukamel D. B., Caprio T., **Ahn R.**, Zheng N. T., Norton S., Quill T., Temkin-Greener H. End of Life Quality of Care Measures for Nursing Homes: Place of Death and Hospice. Journal of Palliative Medicine 15(4):438-446, 2012.

5. Mukamel D. B., Spector W. D., Zinn J., Weimer D. L., **Ahn R.** Changes in Clinical and Hotel Expenditures following Publication of the Nursing Home Compare Report Card. Medical Care 48(10):869-74, 2010.

Refereed Conference Proceedings

1. **Ahn R.**, Garner C.; An empirical validation of random effects and Bayesian meta-analysis models; (1767T). Presented at the 63rd Annual Meeting of The American Society of Human Genetics, Oct. 24, 2013, Boston, MA.
2. **Ahn R.**, Adamson A., Deng X., Gao H., Garner C., Neuhausen S., Fine-scale association mapping of the xMHC-region in celiac disease cases and controls; (681T). Presented at the 61st Annual Meeting of The American Society of Human Genetics, Oct. 13, 2011, Montreal, QC, Canada.

AWARDS

University of California, Irvine:
School of Medicine Travel Grant — 2011 and 2013
Department of Epidemiology First Year Fellowship — 2010-2011

TEACHING EXPERIENCE

University of California, Irvine — 2013
Teaching Assistant
Genetics (undergraduate course): Dr. Rahul Warrior, Dr. Olivier Cinquin, and Dr. Lee Bardwell, Fall 2013
Laboratory in Big Data Analysis (graduate course): Dr. Dana Mukamel, Spring 2013

San Diego State University — 2005-2006
Lecturer
Introductory Macroeconomics (undergraduate course): Spring 2006, Fall 2006
Teaching Assistant
Introductory Macroeconomics (undergraduate course): Dr. Mike Hilmer, Fall 2005

PROFESSIONAL SERVICES, AFFILIATIONS, AND TRAINING

PLOS ONE — 2011-Present
Reviewer

American Society of Human Genetics — 2011-Present
Member
Judge, Annual DNA Day Essay Contest

University of California, Los Angeles, Department of Human Genetics — July 2012
Trainee, 9th Annual Statistical Genetics Short Course

ABSTRACT OF THE DISSERTATION

Genetics of Celiac Disease

By

Richard Sungho Ahn

Doctor of Philosophy in Epidemiology

University of California, Irvine, 2014

Professor Chad Garner, Chair

While most common variants associated with celiac disease have now been identified through genome-wide association studies (GWAS), outstanding questions still exist regarding the validity of previously identified common variants, the presence of common variants associated with celiac disease within the major histocompatibility complex (MHC) region of chromosome 6 that are independent of the high-risk HLA genotypes, and the presence of low-frequency and rare variants associated with celiac disease within previously implicated genomic regions. This dissertation sought to study all of these questions by employing GWAS, fine-mapping methods, meta-analysis methods, imputation, and next-generation sequencing (NGS) of targeted genomic regions.

In the first study of this dissertation, two large-scale celiac disease GWASs were re-analyzed using alternative random-effects meta-analysis models in addition to the fixed-effects approach employed in each GWAS meta-analysis. Implementing a random-effects meta-analysis model did not appreciably increase or decrease the power to detect an association and nearly all of the previously implicated loci were found to be genome-wide significant in the re-analysis. In the second study, a fine-mapping approach of the MHC region that takes into account the effect of the high-risk HLA genotypes was implemented. After adjustment for the high-risk HLA genotypes and the linkage disequilibrium in the MHC region, seven novel loci were found to be associated with celiac

disease. In the third study, targeted NGS-based resequencing was performed on previously implicated genomic regions to test for the presence of low-frequency and rare variants associated with celiac disease. Gene-based collapsing tests revealed that dozens of genes harbor low-frequency and rare variants that are associated with celiac disease, particularly in the MHC region and within non-coding regions of genes. The fourth study implemented a variant imputation method to impute low-frequency and rare variants into a large GWAS dataset to increase the statistical power to detect low-frequency and rare variants. Nearly all of the low-frequency and rare variant associations from the third study were replicated in this fourth study along with novel associations, using both gene-based tests and single-marker association tests.

These studies reveal that there are many more loci that need to be carefully followed-up in larger resequencing studies and functional studies than previously acknowledged by large-scale celiac disease GWAS that do not account for the role of the high-risk HLA genotypes or the low-frequency and rare variants within genomic regions that harbor common variants previously found to be associated with celiac disease.

Chapter 1: Introduction

With the transition from linkage analysis and positional cloning to genome-wide association studies (GWAS) in the post-genome era of the early 2000s, the number of loci known to be associated with numerous, common, complex diseases has grown exponentially. Coupled with microarray genotyping platforms at the single-nucleotide polymorphism (SNP) resolution that have steadily decreased in price while simultaneously increasing in variant density, the agnostic nature of genome-wide scans have quickly grown in popularity amongst both clinical and basic science oriented human genetics researchers. For many common, complex diseases, much of the common variation (minor allele frequency (MAF) ≥ 0.05) conferring low to modest risk of disease (odds ratios 1.1-2) has been identified and replicated via GWAS and meta-analysis of GWASs. In a number of studies, fine-mapping efforts have also been undertaken to identify putatively causal variants that are in linkage disequilibrium (LD) with tag SNPs found on commercially available 'SNP chips' by employing denser genotyping arrays, resequencing of targeted genes, and in the case of autoimmune disorders such as celiac disease, adjustment for the typically high effect size of human leukocyte antigen (HLA) serotypes.

However, while GWAS meta-analysis and fine-mapping methods have dramatically increased the number of known common risk loci, it is not yet clear that widely-used methods of meta-analysis sufficiently adjust for heterogeneity between studies to identify low-frequency or rare variants that may underlie the common variants already uncovered. There are also gaps in the knowledge of causal variants in the 3.6 Mb major histocompatibility complex (MHC) region of chromosome 6 that are not in LD with known alleles of human leukocyte antigen (HLA) serotypes implicated in autoimmune diseases.

This dissertation will empirically re-evaluate prior meta-analyses of celiac disease GWAS by implementing novel meta-analysis models that better incorporate heterogeneity, investigate the use

of fine-mapping methods to identify additional risk loci associated with celiac disease in the MHC region, perform association analysis of low-frequency and rare variants identified through targeted resequencing of coding regions of celiac disease associated genes identified in previous GWAS, and finally, utilize an imputation method to impute rare and low-frequency SNPs into a large-scale GWAS dataset to identify putatively causal, low-frequency and rare variants genome-wide and in the MHC region that may have gone undetected in previous celiac disease GWASs and meta-analyses GWAS. This chapter will provide a brief introduction to GWAS, meta-analysis, and celiac disease as well as providing a summary of the specific aims and hypotheses of this dissertation.

1.1. Genome-wide Association Studies

As the Human Genome Project approached completion[1], researchers began to transition away from linkage analysis based studies. While linkage analysis has yielded remarkable results in the identification and mapping of highly penetrant genes linked to susceptibility for hundreds of Mendelian disorders throughout the 1980s through the 1990s, including monogenic disorders such as cystic fibrosis and Huntington's disease[2–4]. In general, the risk alleles in the susceptibility genes identified by linkage analysis tended to be non-synonymous, or amino acid altering, and of low-frequency[5]. However, while these approaches have found significant results in many studies of common, low-penetrance diseases, very few have led to convincing replication studies, with one estimate suggesting that perhaps less than 10% of non-Mendelian disease loci may be identified through linkage analysis[6–8].

Linkage-analysis based studies of complex disorders tended to yield more modest and non-replicable results relative to Mendelian disorders. The high costs associated with performing whole-genome linkage analysis studies involving the collection of large family pedigrees spurred researchers to re-examine the potentials of performing association studies on candidate genes or across the

whole genome[8]. With the completion of the Human Genome Project[9] and the advent of relatively inexpensive microarray-based genotyping machines that could capture hundreds of thousands to millions of polymorphisms for thousands of individuals in a short span of time, adoption of genome-wide association studies (GWAS) began in earnest around 2005[10] with the number of new GWAS for a variety of complex disorders growing exponentially for the next few years. In a relatively short span of time, GWAS have uncovered many common variants associated with complex diseases and traits such as age-related macular degeneration, type 2 diabetes, Crohn's disease, systemic lupus erythematosus, and cholesterol levels[10–14].

Non-Mendelian diseases are typically governed by risk-susceptibility alleles across several genes and patterns of inheritance within families tend to be more complicated than in Mendelian diseases[15]. This is in addition to the environmental and other random factors that may or may not be accounted for. Underlying GWAS, are two competing hypotheses to explain the role of variants in the etiology of complex diseases. The first is the 'common disease-common variant' hypothesis (CDCV). According to CDCV, complex disease may be largely explained by a set of common variants of low-penetrance, with each variant explaining some small percentage of the population risk[16]. The second is the 'common disease-rare variant hypothesis' (CDRV). According to CDRV, a large proportion of the population risk may be attributable to the effect of a few variants of low MAF and of high-penetrance[17]. Common variants tend to have modest ORs between 1.1 and 2, with very few variants having an OR equal to or greater than 2, while most rare or low-frequency variants are thought to have ORs greater than 2. Many researchers have also adopted a hybrid model in which common and rare variants together explain the risk of developing a complex disease.

While CDCV posits that common variants may be causal, most of the common variants that have been identified in GWASs thus far are probably not causal because they may be in LD with some underlying causal variant that remains undetected in GWAS because the variant was not

captured on commercially available SNP arrays[18,19]. In many instances, the best evidence for a loci being identified as either causal or in linkage disequilibrium with a causal variant is the *p*-value of the association and to achieve genome-wide significance, the *p*-value must be extremely low, with common acceptance of a *p*-value of $5x10^{-8}$ or lower as the threshold[20–22]. To detect smaller and smaller odds ratios, researchers need to genotype larger numbers of samples to have sufficient power to detect a genome-wide significant association. However, the marginal gains in utility have been rapidly diminishing[5].

Next-generation sequencing (NGS) studies now allow researchers to better evaluate the distribution of allele frequencies, both common and rare variants that are likely to be associated with common, complex diseases[22]. Common variants, particularly in the intergenic regions, may act as effect modifiers for the rare variants or be in linkage-disequilibrium with the rare variants and act as markers for candidate loci that are to be resequenced. There is strong evidence for the latter view and this evidence has had a significant impact on the design of denser, custom microarrays and in the design of statistical tests to detect rare variants[15,23]. Recent resequencing-based association studies of inflammatory bowel disease, multiple sclerosis, and age-related macular degeneration[24–27] provide evidence that resequencing studies may provide sufficient power to discover high-risk rare variants under CDRV[28].

1.2. Meta-analysis of GWAS

While individual GWAS led to the discovery of many novel polymorphisms associated with a host of complex diseases, replication studies to provide robust evidence of association have now become de rigueur. However, even for common variants, sufficiently powered replication studies require tens or hundreds of thousands of samples to be genotyped to minimize the number of false-positives, a requirement that any individual investigator may find infeasible because of the severe

cost[29–32]. These constraints have led to the widespread and rapid adoption of meta-analysis of GWAS beginning around 2007 and the formation of large consortia[33]. As a result, many of the risk variants discovered and replicated in the last half-decade are the fruit of meta-analyses of GWASs, with several hundred meta-analyses of GWAS having been published at the present time[34,35]. These studies based on consortium data have provided strong evidence for small effect sizes (in terms of ORs) that range from 1.1 to less than 2[18,22,36,37].

Within the biomedical sciences, the method of combining p-values—and more specifically, Fisher's method—was the most popular meta-analysis method. It was largely abandoned around the 1980s because of limitations that include the inability to estimate a summary effect across all studies and the inability to deal with heterogeneity between studies[38]. The most popular approach—and the most powerful—is fixed-effects meta-analysis[39]. Fixed-effects meta-analysis effectively yields results that are similar to the Cochran-Mantel-Haenszel method[35,38] for obtaining a pooled odds ratio and p-value in 2x2x*k* contingency table analysis or the weighted Z-scores method that yields results similar to fixed-effects meta-analysis when the fixed-effects weight that is used is the inverse of the variance for each study, the optimal method of weighting a fixed-effects model[40,41]. The random-effects model is the preferred model when between-study heterogeneity is detected because it allows for greater generalizability of association results discovered in GWAS meta-analysis across different population groups. The random-effects model incorporates the between-study heterogeneity into the effect estimator weight[42]. While false-positive rates are higher when using the random-effects model with few data sets (as is the case in a discovery stage), it also has a much lower false-positive rate as the number of data sets increases, lower by several magnitudes at times. This tendency makes random-effects desirable in meta-analyses aimed at replications[43].

1.3. Celiac disease

Celiac disease, or coeliac sprue, is a chronic, autoimmune, heritable, enteropathy of the small intestine that is triggered by ingestion of gluten peptides found in wheat, barley, or rye products. First diagnosed in modern times in the late nineteenth century by Samuel Gee, typical symptoms include bloating, abdominal pain, chronic diarrhea, and failure to thrive in affected children[44]. Currently available therapeutic treatment consists solely of strict dietary exclusion of gluten and supplements to treat vitamin and mineral deficiencies that result from malabsorption of nutrients in the small intestine[45]. Left untreated, these vitamin and mineral deficiencies may lead to anemia, osteoporosis, and neurological disorders[46,47]. While over 90% of patients respond very well to a gluten exclusion diet, around 5% of patients, particularly those patients that developed symptoms past age 50, do not respond to treatment and are then diagnosed with refractory celiac disease[48]. The estimated prevalence of disease is between 0.5 and 1.26% amongst Caucasians in Europe and USA[49]. The global prevalence of disease is much lower with an estimated prevalence around 0.03%[50,51]. Even after adjustment for advancements in screening technology and historic underestimation of the occurrence rate of celiac disease, the incidence of celiac disease has been shown to be increasing over time[52]. Celiac disease is more prevalent among women than men below the age of 60 and it is thought that there may be some loci that are influenced by gender and by hormone levels. Interestingly, amongst cases greater than 60, celiac disease is more prevalent among males[53]. While celiac disease may be diagnosed at any age, most cases are diagnosed in early childhood or around the fourth or fifth decade of life[51,54]. Perhaps the most troubling complication occurs primarily in patients diagnosed with celiac disease after the age of 50 and do not have a remission of symptoms after being placed on a gluten-free diet. These patients are diagnosed as having refractory celiac disease and are at a significantly higher risk for developing enteropathy-associated T-cell lymphoma[55].

It has been well established in the last three decades that individuals that carry the alleles for the HLA-DQ2 or HLA-DQ8 haplotypes that encode for MHC class II heterodimeric molecules on the surface of antigen-presenting cells are more susceptible to developing celiac disease than non-carriers. Most cases carry either one or both of the DQ2 and DQ8 haplotypes, with the majority being heterozygous carriers of the DQ2 haplotype. However, while DQ2 and DQ8 haplotypes are widely considered to be necessary in the development of celiac disease, it is not considered to be sufficient for diagnosis and prediction of the onset of disease as some 40% of healthy individuals in the highest-risk Western European population are carriers of alleles for either the DQ2 or DQ8 haplotypes[56–59]. Highly sensitive and specific serological tests for the presence of IgA antibodies versus tissue transglutaminase are the current gold standard for diagnosis of celiac disease along with tissue biopsy of the small intestine and positive response to exclusion of dietary gluten[56,60]. According to the diagnosis algorithm of the ESPGHAN criteria, a patient is positively diagnosed with celiac disease when histopathological analysis reveals hyper plastic villous atrophy of the small-intestine and a remission of symptoms after the patient has been placed on a gluten-free diet[56,61,62]. Besides the well-established environmental risk factor of gluten peptides and the genetic risk factor of HLA-DQ2 and HLA-DQ8 molecules, there has also been speculation that adenovirus and rotavirus infection after birth may increase the risk of celiac disease[63,64].

Genetic risk factors other than the genotypes encoding for HLA-DQ2 and HLA-DQ8 have been discovered in several linkage analysis studies[65–67]. The first GWAS (and GWAS meta-analysis) of celiac disease by van Heel et al.[68] identified variants in the *IL2* and *IL21* genes, genes critical in the expression of T-cell cytokines. A follow-up study by Hunt et al.[69], replicated the results from van Heel et al. and also identified six new loci that regulate the adaptive immune response in celiac disease. Trynka et al.[70] performed a follow-up study to the Hunt et al. study with four more study populations and were able to identify two novel loci associated with celiac disease, *OLIG3-*

TNFAIP3 and *REL*. Garner et al.[71] also performed a follow-up to the Hunt et al. study with samples drawn from a US population and was able successfully replicate five of the eight regions identified in the previous study, including *IL2* and *IL21* and was able to provide evidence for a new loci associated with celiac disease, *ITGA4*. In 2010, Dubois et al.[72] performed a large-scale meta-analysis GWAS involving twelve population groups and were able to identify thirteen more novel loci with genome-wide significant evidence with another thirteen novel loci with suggestive evidence. Using a custom, dense genotyping platform, Trynka et al.[73] identified another thirteen novel celiac disease loci and performed fine-mapping leading to the discovery of multiple independent variants at about a third of the novel loci, providing evidence that a combination of common and rare variants play a role in increasing risk for developing celiac disease. Another fine-mapping study was performed by Ahn et al.[74] using genotype data from a North American celiac disease GWAS that was able to identify four new loci in the extended MHC region that are independent of the HLA genotypes. Most recently, Hunt et al.[75] conducted a large-scale exon sequencing based study of 25 genes previously implicated in celiac disease GWAS with over 40,000 samples and found that rare variants in coding regions explain very little heritability of disease (approximately 3%) and claim that there is little evidence to support the implementation of large-scale whole-exome sequencing studies in autoimmune diseases such as celiac diseases although this study does not provide evidence against implementing whole-genome or targeted resequencing studies that include exonic, intronic, and intergenic regions. Including all studies to date, all HLA and non-HLA loci have been estimated to explain approximately 60% of the genetic variance of celiac disease[72,73,75], indicating that there is much heritability left to be explained.

1.4. Hypotheses and experiments

Hypothesis 1: Utilizing novel random-effects meta-analysis methods will increase the power to detect common variants at previously identified loci with genome-wide significance.

Experiment 1 (Chapter 3): Two large-scale celiac disease GWAS that used a fixed-effects equivalent meta-analysis approach[72,73]. This chapter investigates the change in statistical significance when two alternative random-effects were utilized to account for between-study heterogeneity. Standard fixed-effects and random-effects models were also utilized for comparison purposes.

Hypothesis 2: There are novel variants associated with celiac disease in the extended MHC region that are independent of the high-risk HLA genotypes.

Experiment 2 (Chapter 4): While the high-risk HLA genotypes within the classical MHC region are necessary for disease development, it has been difficult to determine if there are any HLA-independent loci within the MHC region due to the complexity of the region. A fine-mapping approach was implemented to determine if there are variants associated with celiac disease in the extended MHC region independent of the high-risk genotypes for HLA-DQ2 and HLA-DQ8.

Hypothesis 3: There are rare and low-frequency variants associated with celiac disease within genes previously found to be associated with celiac disease.

Experiment 3 (Chapter 5): A small-scale (approximately 500 samples) targeted NGS-based resequencing study was conducted to determine if there are rare and low-frequency variants that are associated with celiac disease. Gene-based collapsing tests were used to test for association as a single-variant test will be underpowered to detect an association.

Hypothesis 4: Rare and low-frequency variants imputed forward into a previous large-scale GWAS dataset will be well powered to provide statistically significant evidence of association with celiac disease under gene-based and single-variant tests.

Experiment 4 (Chapter 6): While large-scale GWAS using dense microarray genotyping platforms have been performed, the vast majority of rare and low-frequency variants have not been tested for association with celiac disease. A variant imputation method was utilized to impute forward rare and low-frequency variants into a large-scale GWAS dataset (~24,000 samples) and both gene-based and single-variant tests were performed to test for association.

Chapter 2: Methods and Materials

This chapter describes the subjects, the investigators and institutions that collected the data, how the genotyping of samples was conducted, and the phenotyping of samples. Several institutions in the United States of America, the European Union, as well as India were involved in the collection of samples. Genotyping was performed in the US, The Netherlands, and the UK. All samples were phenotyped according to similar protocols.

2.1. Subjects

Subjects for association analysis of the extended MHC region in celiac disease and association of rare and low-frequency variants in the 12 previously identified regions and the MHC region with celiac disease

2300 cases and controls were obtained from the North American Celiac Disease Genetic consortium[74]. This consortium is led by Dr. Susan Neuhausen at City of Hope (COH) in Duarte, CA, Dr. Chad Garner at the University of California, Irvine (UCI), Dr. Joseph Murray at the Mayo Clinic in Rochester, MN, Dr. Alessio Fasano at the University of Maryland, and Dr. Peter Green at Columbia University, in NYC, NY. The institutional review board (IRB) of each respective institution has approved the IRB protocols submitted at each institution to collect, code, and analyze samples. Every IRB protocol included a written informed consent form signed by all subjects. Of the 2300 total subjects, 1764 are cases and 536 are controls. Of the 1764 cases, 532 were collected at the University of Utah and UCI, 743 were collected at the Mayo Clinic, 423 were collected at the University of Maryland, and 66 were collected at Columbia University. Of the 536 controls, 177 were collected at COH and 359 were collected at the Mayo Clinic (Table 2.1). As celiac disease predominantly affects Caucasian populations, all of the 2300 cases and controls are Caucasian and unrelated. All subjects had blood samples collected for serological testing and DNA extraction. Of

the 1068 subjects collected at COH, 247 cases, 234 controls, and 26 phenotypically ambiguous samples were resequenced.

Table 2.1. Number of cases and controls by institution from Ahn et al.[74]

Institution	Case	Control
COH	532	177
Mayo Clinic	743	359
UMaryland	423	-
Columbia	66	-

Subjects for analysis of meta-analysis methods for genome-wide association studies and application of meta-analysis methods in association studies of celiac disease

All samples were collected for previously reported multi-stage, multi-country GWASs by the European Celiac Disease Consortium and were generously provided by Dr. David van Heel at the Blizard Institute of Cell and Molecular Science, Barts and the London School of Medicine and Dentistry, Queen Mary University of London, London, UK (Tables 2.2, 2.3, and 2.4)[72,76]. The first UK cohort cases were collected at seven hospitals in the UK: Barts and the London, London; Hammersmith Hospital, London; Leeds University Hospitals, Leeds; Llandough Hospital, Cardiff; Sheffield University Hospitals, Sheffield; Derbyshire Royal Infirmary, Derby; John Radcliffe Hospital, Oxford[68]. The second UK cohort cases were collected at the same hospitals as the first UK cohort, with the exception of 374 cases and 176 controls that were collected by the Celiac UK recruitment effort. Any cases in the first UK cohort showing genetic relatedness to cases in the second UK cohort were removed. All UK cases were matched to a control from either the 1958 British Birth Cohort or the National Service Cohort. Written informed consent was obtained for all subjects and ethics approval was provided by the Oxfordshire Research Ethics Committee B as well as local ethics committees[69]. All subjects were Caucasian of northern European ethnic origin and unrelated. The first UK cohort contained 737 cases and 2596 controls while the second UK cohort contained 1849 cases and 4936 controls. Another 6209 cases from the Celiac UK recruitment effort

and 742 controls from the 1958 British Birth Cohort and the National Service Cohort were added to the first and second UK cohorts for a dense-genotyping based fine-mapping study[76].

The cases for the first and second Finnish cohorts were collected at the University of Tampere[72,77]. Matched population controls were collected through the Finrisk cohort and the Health 2000 cohort. All subjects provided a written informed consent and ethics approval was provided by the ethics committees of the University of Tampere, Helsinki University Hospital, and the Finnish National Public Health Institute. All subjects were ethnically Finnish and unrelated. The first Finnish cohort contained 647 cases and 1829 controls while the second Finnish cohort contained 259 cases and 653 controls.

The cases and controls for the first Italian cohort were collected at the Centro per la prevenzione e diagnosi della malattia celiaca, Fondazione IRCCS Ospedale Maggiore Policlinico in Milan, Italy[78]. Cases and controls for the second Italian cohort were collected at the Pediatric Department of the Sapienza University of Rome[79] in Rome, Italy. Written informed consent forms were obtained from all subjects and study approval was given by the ethics committee of the Fondazione IRCCS Ospedale Maggiore Policlinico. All subjects were ethnically Italian. The first Italian cohort contained 497 cases and 543 controls while the second Italian cohort contained 1010 cases and 804 controls.

Dutch cases and controls were collected at the University Medical Center in Utrecht, The Netherlands[68,80]. All subjects had provided written informed consent forms and approval was obtained from the medical ethics committee of the University Medical Center in Utrecht. All subjects were ethnically Dutch and unrelated. The Dutch cohort contained 803 cases and 846 controls. A further 320 cases and 301 controls were obtained under the same protocol for a dense-genotyping based fine-mapping study[76].

Cases from the USA were collected at the Mayo Clinic, Rochester MN and at UCI, Irvine CA[71]. All subjects had signed written informed consent forms and approval for the study was provided by the IRBs at the Mayo Clinic and UCI. All subjects were Caucasian and unrelated. Age, sex, and ethnicity matched population controls were obtained from the COH and the Mayo Clinic. The US cohort contained 973 cases and 555 controls.

Irish cases were collected at St. James' Hospital and the Adelaide and Meath Hospitals in Dublin, Ireland, and at University College Hospital, Galway, Ireland. Study approval by the Institutional Ethics Committee of St. James' Hospital and local approval was obtained along with signed written informed consent forms for all subjects. All subjects were of Caucasian, northern European origin and unrelated. The Irish cohort contained 597 cases and 1456 controls[68].

Polish cases were collected in hospital clinics throughout Poland while controls were collected from the Children's Memorial Health Institute in Warsaw, Poland. Signed written informed consent forms were obtained for all subjects and approval were obtained by the local ethics committee. The Polish cohort contained 564 cases and 716 controls[72].

Hungarian cases were collected from children's clinics in Budapest and Debrecen and population matched controls were obtained from a previous epidemiological study. Signed written informed consent forms and the approval of local ethics committees were obtained. All subjects were ethnically Hungarian and unrelated. The Hungarian cohort contained 965 cases and 1067 controls[72,81].

Spanish cases were collected from Madrid area hospitals while controls were obtained from hospital employees and blood donors. Signed written informed consent forms and the approval of the ethics committee at Hospital Clinico San Carlos were obtained. All subjects are Caucasian in origin and unrelated[82].

The Indian cohort (only used in the dense-genotyping based fine-mapping study) was collected from the Punjab region of India[76] with 229 cases and 391 controls. Signed written informed consent forms and the approval of local ethics committees were obtained. All subjects were ethnically Indian and unrelated; this is the only non-European cohort collected.

Table 2.2. Number of cases and controls by country for stage 1: GWAS of Dubois et al.[72]

Country	Case	Control
UK	737	2596
UK2	1849	4936
Finland	647	1829
The Netherlands	803	960
Italy	497	580

Table 2.3. Number of cases and controls by country for stage 2: follow-up of Dubois et al.[72]

Country	Case	Control
USA	973	555
Hungary	965	1067
Ireland	597	1456
Poland	564	716
Spain	550	433
Italy	1010	804
Finland	259	653

Table 2.4. Number of cases and controls by country of Trynka et al.[76]

Country	Case	Control
UK	7728	8274
The Netherlands	1123	1147
Poland	505	533
Spain-CEGEC	545	308
Spain-Madrid	537	320
Italy-Rome, Milan, Naples	1374	1255
India-Punjab	229	391

2.2. Phenotyping

Phenotyping of subjects for association analysis of the extended MHC region in celiac disease and association of rare and low-frequency variants in the 12 previously identified regions and the MHC region with celiac disease

Each of the four institutions that collected samples for the North American Celiac Disease Genetic Consortium used the same serology tests and the same diagnostic criteria for positively identifying a celiac disease case[74]. To be defined as a case, each subject had to meet two of three diagnostic criteria: 1) be positive for IgA EMA and IgA tTG antibodies; 2) or exhibit a small intestinal biopsy that is indicative of celiac disease; and e) demonstrate either a clinical or histological improvement on a gluten-free regimen. Most of the cases met all three criteria; a small proportion of subjects that were collected before the development of celiac-specific antibodies did not have a positive serology while another small proportion of subjects did not report a biopsy. Self-reported cases or subjects with biopsies indicative of only a minor infiltration of intraepithelial lymphocytes were not considered cases (Figure 2.1).

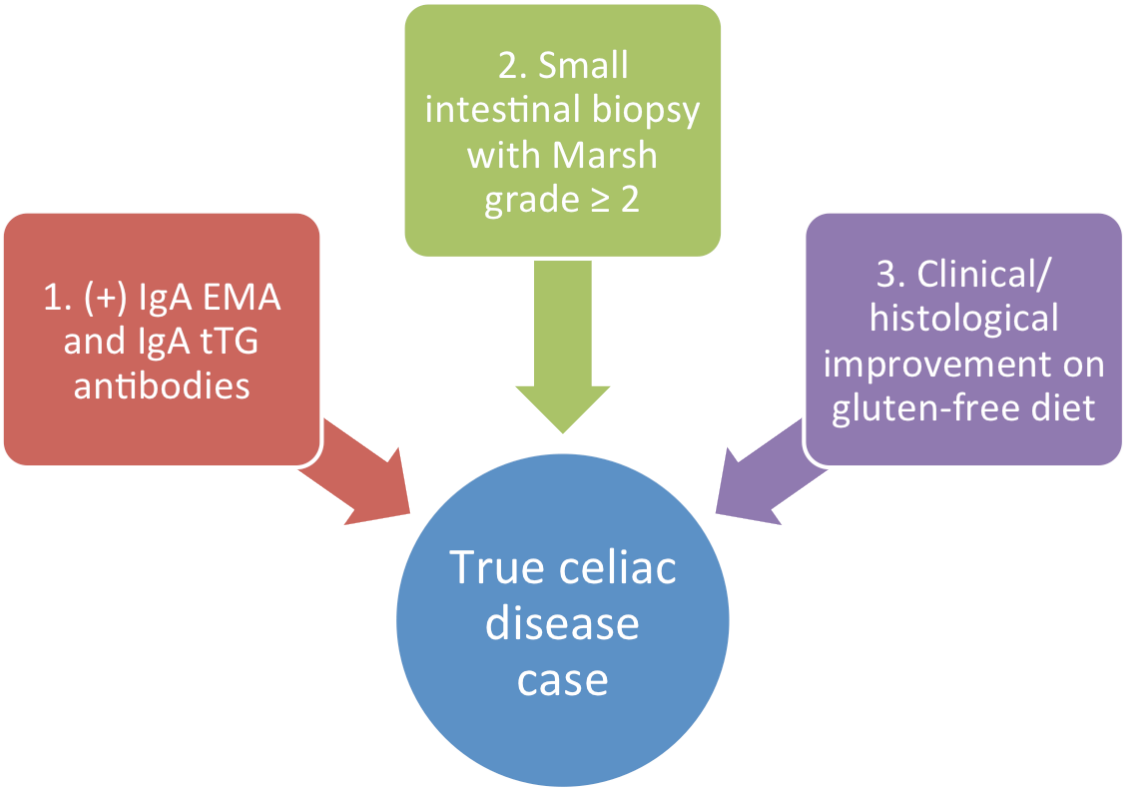


Figure 2.1. Meeting any two of the three criteria for diagnosing celiac disease indicates a true positive case.

Phenotyping of subjects for analysis of meta-analysis methods for genome-wide association studies and application of meta-analysis methods in association studies of celiac disease

Nearly all cases were diagnosed according to the criteria set forth in the revised European Society for Paediatric Gastroenterology, Hepatology, and Nutrition (ESPGAN)[83]: these criteria include a positive serology test for celiac-specific antibodies, clinical, and histopathological criteria. All cases had a small intestinal biopsy indicating the presence of celiac disease. Individual collection centers had variations in biopsy criteria. UK, Dutch, Polish, Italian, Hungarian, and Indian cases were required to exhibit at least a Marsh grade III biopsy. Cases from Spain were required to have at least a Marsh grade II biopsy.

2.3. Genotyping

Genotyping of subjects for association analysis of the extended MHC region in celiac disease and association of rare and low-frequency variants in the 12 previously identified regions and the MHC region with celiac disease

All 2300 samples were genotyped at Center for Inherited Disease Research at Johns Hopkins on the Illumina 660 W Quad platform. Samples and SNPs were excluded if data was missing for more than 2% of either the sample or the SNP. SNPs with MAF < 0.03 or failing the Hardy-Weinberg Equilibrium (HWE) with $p < 1x10^{-5}$ were also excluded. Samples were also tested for genetic relatedness and excluded if genetically related. Reported sex was also tested for and if the reported sex did not match the genetic sex, the sample was excluded. Samples were also tested for population stratification via multi-dimensional scaling (MDS)[84] and subsequent cluster analysis to remove outlier individuals. A total of 114 samples were excluded leaving 517 controls and 1668 cases that passed all QC criteria. 1898 SNPs from the xMHC (chromosome 6p; between positions 26000508 and 33544122) were included from the entire GWAS panel. Sanger sequencing was

performed for 95 samples to determine the HLA-DQA1 and DQB1 alleles to determine controls for HLA genotyping using a PCR method developed at COH. The remaining samples were genotyped by either a PCR-based HLA-DQ typing method developed at COH[74] or by a tag SNP method that utilizes six SNPs to predict the four HLA-DQ types known to be associated with celiac disease.

508 samples from the COH were resequenced on the Illumina Genome Analyzer IIx platform at an average read depth of 35x after targeted enrichment of the 12 previously identified regions and the 7.6 Mb xMHC region of chromosome 6p. The sequence libraries were enriched with either the Agilent or NimbleGen bait platform. A custom pipeline was implemented using BWA, SAMtools, Picard, and the GATK packages to align the paired-end reads to the reference genome and call SNP[85–87].

Genotyping of subjects for analysis of meta-analysis methods for genome-wide association studies and application of meta-analysis methods in association studies of celiac disease

All GWAS genotyping (first stage) and follow-up genotyping (second stage) (Figure 2.2) was performed in labs in London, Hinxton, and Groningen. The UK1 cases were genotyped on the Illumina Hap300v1-1 platform (~300K SNPs) while the UK1 controls were genotyped on the Illumina Hap550-2v3. The UK2 cases, Finland1 cases, the Dutch cases and controls, and the Italian cases and controls were all genotyped on the Illumina 670-QuadCustom_v1. The UK2 controls were genotyped on the Illumina 1.2M-DuoCustom_v1 and the Finland1 controls were genotyped on the Illumina 610-Quad. All follow-up genotyping of the 131 SNPs for the USA, Hungary, Ireland, Poland, Spain, Italy2 cases and controls (and the Finland2 cases) was performed on the Illumina GoldenGate BeadXpress platform. Finally, the Finland2 controls were genotyped using the Illumina 610-Quad platform.

Dense genotyping of the 183 non-HLA loci (~196K SNPs) was performed for all population cohorts (UK, Dutch, Polish, Spanish, Italian, Indian) using the custom Illumina Immunochip platform in labs in London, UK, Hinxton, UK, Groningen, The Netherlands, and Charlottesville, USA.

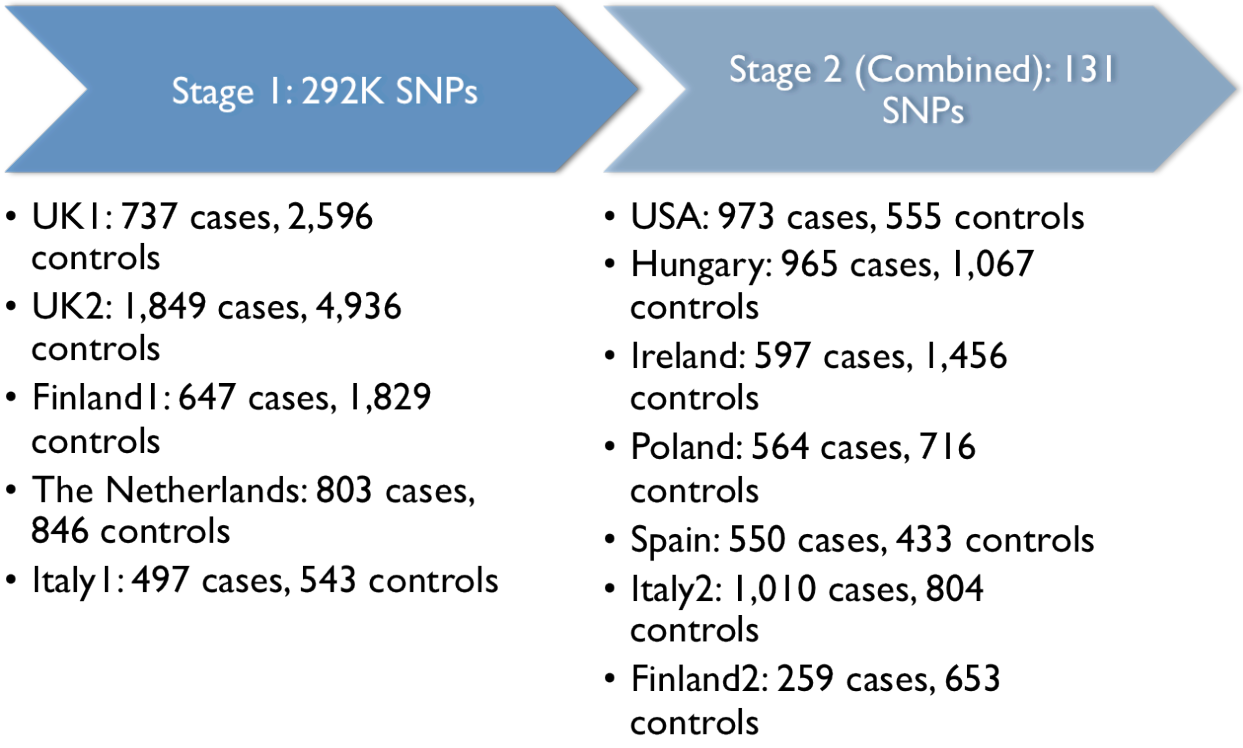


Figure 2.2. First stage GWAS with a second stage follow-up with only the SNPs that passed a set threshold in the first stage. Adapted from Dubois et al.[72]

Chapter 3: Meta-analysis Methods for Genome-wide Association Studies

Genome-wide association study (GWAS) meta-analysis (MA) is now routinely used to combine either individual-level data or summary statistics from multiple GWASs to increase statistical power to detect the small effect sizes of common alleles and to decrease the likelihood of observing false-positive associations. The cost to perform a large-scale GWAS MA with sample sizes ranging from the tens of thousands to hundreds of thousands is reduced by orders of magnitude because new samples do not have to be genotyped. As a result of this reduction in experimental cost, hundreds of GWAS MAs have been performed over the last half-decade and have significantly increased the quantity of risk loci discovered and replicated for a number of different phenotypes. There are now several approaches to GWAS MA, with the most widely implemented approach being fixed-effects MA because it is the least conservative and most statistically powerful approach for most GWAS MAs. However, fixed-effects MA is not always ideal because it ignores the potential heterogeneity or between-study variance that may exist between studies by assuming that the effect of a risk allele is homogeneous across all of the studies in the MA. In this study, two GWAS meta-analyses of celiac disease were reanalyzed: 1) A GWAS MA that includes 9,451 celiac disease cases and 16,434 controls from 12 collections and 2) a GWAS MA using a custom dense genotyping platform to capture variants across a greater allelic spectrum in 12,041 cases and 12,228 controls from 7 collections. The purpose of this study was to determine if the results from a random effects MA that accounts for between-study heterogeneity will differ significantly from the results that were originally published. This study presents evidence that a SNP at a locus in chromosome 1 (*RUNX3*) that was previously reported to show genome-wide significance ($p < 5x10^{-8}$) in one of the previous GWAS MA was not genome-wide significant in either the discovery stage or the combined stage when the between-study heterogeneity was adjusted for by the random effects MA approaches and

highlights the need to carefully investigate between-study heterogeneity and the implementation of GWAS MA models that control for heterogeneity.

3.1. Introduction

For several decades, MA has been implemented in many disparate fields to pool together the estimated summary statistics from independently conducted studies of a given trait or disease and increase statistical power to detect small effect sizes that would otherwise be undetected in each study of the MA[40,42,88]. In the last half-decade, MA has been heavily employed within the context of genome-wide association studies of human diseases. GWAS MA was adopted early on by human disease investigators because it provides increased power to detect associations of small effect size for common diseases and complex traits by combining either the summary statistics or individual-level data from previously conducted studies, thereby lowering the cost to detect new associations by allowing investigators to avoid having to perform expensive de novo collection and genotyping of the thousands of cases and controls required for sufficient power. As genotypic data are readily available through several consortia and public data repositories, a large number of MAs of many common diseases and traits[11,13,88–92] have already been performed and have identified new associations that could not be detected in any one of the component studies. Uncertainty regarding the validity of the results from early GWAS MAs[35,93–96] have centered on the effects of between-study heterogeneity on the results of a GWAS MA and how to properly adjust for between-study heterogeneity arising from the aggregation of multiple independent studies[38].

In previous GWASs of celiac disease, discovery and replication studies were combined by implementing a fixed-effects (FE) MA framework from samples representing diverse population groups[68,69,72] ranging from the UK, the Netherlands, Italy, Spain, Poland, Hungary, Finland, Ireland and the USA. The MA framework that was implemented was a FE equivalent model in the form of

the Cochran-Mantel-Haenzel (FE-CMH) test[40,97] that led to the discovery and replication of 13 novel celiac disease loci in Dubois et al.[72]. Trynka et al.[76] performed a dense genotyping study of previously identified celiac disease regions using a custom genotyping platform with a higher density of SNPs that have a wider allele frequency spectrum than previously analyzed. This study included seven independent sample collections (six from Europe and one from India) and also employed a FE equivalent model based on conditional logistic regression (FE-CLR)[98,99] that included a categorical variable for the collection ethnicity. An FE equivalent model such as FE-CLR pools data from each study or collection prior to the estimation of the effect size and variance and does not explicitly weight the overall effect size and p-value by the within-study or between-study variance. However, MA models that explicitly combine effect sizes such as the FE and random-effects (RE) models, estimate the effect size and variance for each study independently before combining and weighting the summary statistics from each study to estimate the overall effect size and p-values.

Using the individual-level data from Dubois et al.[72] and Trynka et al. [76] to estimate individual study effect sizes and within-study variance, the performance, in terms of p-values, of MA implementations that can explicitly account for the observed between-study heterogeneity relative to the performance of FE equivalent methods such as FE-CMH or FE-CLR was compared. Finally, the relative performance of several FE and RE models, including two newly developed random-effects based models[100,101], is compared in the presence of between-study heterogeneity.

3.2. Methods

Measures of heterogeneity

Several measures have been developed to express between-study heterogeneity, with the two most widely adopted measures of between-study heterogeneity being Cochran's Q statistic and the I^2 statistic[99,102], with Q and I^2 as follows,

$$Q = \sum_i w_i (X_i - \mu)^2$$

$$I^2 = \begin{cases} \frac{Q-(k-1)}{Q} \times 100\%, & for\ Q > (k-1) \\ 0, & for\ Q \leq (k-1) \end{cases}$$

where w_i is the weight for a given study *i*, X_i is the effect size estimate for study *i*, μ is the mean effect size estimate across all *i* studies, and *k* is the total number of studies in the MA. As Cochran's *Q* is the sum over *i* studies of the product of the inverse variance for each study *i* and the squared difference of the observed effect size of study *i* and the expected effect size of study *i*, Cochran's *Q* follows a χ^2 distribution with *k*-1 degrees of freedom with the I^2 statistic being based on Cochran's *Q* statistic. Whereas Cochran's *Q* is a statistical test to determine if heterogeneity is present or not, the I^2 statistic expresses between-study heterogeneity as a percentage of the effect size variance due to between-study variance, the τ^2 (see below for detail on estimation of τ^2). The distribution of I^2 ranges from 0% to 100% where an I^2 value of 0% to 40% indicates little evidence of between-study heterogeneity, 40% to 75% indicates evidence of moderate to strong between-study heterogeneity, and an $I^2 > 75\%$ indicates very strong evidence of between-study heterogeneity[102]. The I^2 statistic truncates to zero when $Q \leq k$-1. To estimate the standard error and confidence interval for the I^2 statistic, the H^2 index is calculated[102],

$$H^2 = \frac{Q}{k-1}.$$

The H^2 index has the following relationship with the I^2 statistic,

$$I^2 = \frac{H^2-1}{H^2} \times 100\%.$$

Then, the construction of the confidence interval for *H* is as follows,

$$e^{ln(H) \pm \left| z\alpha_{/2} \right| SE[ln(H)]}.$$

The natural logarithm of *H* is taken to assume a standard normal distribution for the confidence interval and the standard error of ln(*H*), SE[ln(*H*)], is calculated

$$SE[ln(H)] = \begin{cases} \frac{1}{2}\frac{ln(Q)-ln(k-1)}{\sqrt{2Q}-\sqrt{2k-3}} & if\ Q > k \\ \sqrt{\frac{1}{2(k-2)}\left(1-\frac{1}{3(k-2)^2}\right)} & if\ Q \leq k \end{cases}$$

The standard error and confidence interval for I^2 can then be obtained by applying the relationship that H^2 has with I^2.

While both Cochran's Q statistic and the I^2 statistic have low statistical power to detect between-study heterogeneity when the number of studies is low and excessive power when the number of studies is very high, the I^2 statistic has become the preferred statistic to quantify between-study heterogeneity because it is the percentage of variation at a given locus—across all studies in an MA—explained by the between-study variance[94,103] and is thus much easier to interpret.

Fixed-effects

For decades, Fisher's method of combining p-values was the most commonly applied MA method in the biomedical sciences. However, it was largely abandoned because of the method's severe limitations, including an inability to estimate an overall effect size estimate across all studies, an inability to estimate between-study heterogeneity, and a higher false-positive rate of association when the directionality of effect estimates was not consistent across all studies in the MA[38]. The method of weighted Z-scores was an improvement upon Fisher's method of combining p-values, as it allowed for individual weights for each study (as compared to the uniform, and most likely suboptimal, weighting of Fisher's method) and took the directionality of effects into account when estimating an overall Z-score[104].

Under the very ideal assumption of no between-study heterogeneity (i.e. homogeneity), an investigator can maximize the statistical power of pooling together the summary statistics from all available studies by performing a MA under the fixed-effects (FE) model[35,38]. The FE model assumes that effect size estimates, X_i, are homogeneous and normally distributed across each study *i*. The FE

model estimator weights the effect size estimate for each study *i* by the inverse of the within-study variance (i.e. the sampling error) for each study *i*,

$$\overline{X} = \frac{\sum_i W_i X_i}{\sum_i W_i},$$

where $W_i = V_i^{-1}$ and V_i is the standard error of X_i squared. Then, the test statistic for the FE estimator, Z_{FE} is

$$Z_{FE} = \frac{\overline{X}}{SE(\overline{X})}.$$

$\overline{X}$ is assumed to be normally distributed, so the p-values for Z_{FE} may be obtained from the CDF of the standard normal distribution. Han et al.[100] demonstrated that the p-values of the FE estimator are the equivalent of the p-values obtained by taking the weighted sum of *Z* scores of each study *i*,

$$Z_{WS} = \frac{\sum_i \sqrt{N_i p_i (1-p_i)_i}\, Z_i}{\sqrt{\sum_i N_i p_i (1-p_i)}},$$

where N_i is the sample size of study *i* and p_i is the minor allele frequency (MAF) of a given marker from study *i*.

Random-effects

The within-study variance estimates the sampling error within a particular study, while the variance in effect size may exist between the studies in a MA and is known as the between-study variance or heterogeneity, τ^2. In practice, the most commonly used estimator for τ^2 is the DerSimonian and Laird estimator[42],

$$\hat{\tau}^2 = \begin{cases} \frac{Q-(k-1)}{\sum_i w_i - \frac{\sum_i w_i^2}{\sum_i w_i}}, for\ Q > (k-1) \\ 0, for\ Q \leq (k-1) \end{cases}$$

If between-study heterogeneity in effect-size estimates is suspected or estimated across studies in a MA, the random-effects (RE) model is thought to be more appropriate as the RE model assumes heterogeneity to exist under the null hypothesis[42,100] and incorporates the $\widehat{\tau^2}$ into the weighting. Hybrid approaches combining FE and RE models to optimize p-values and lower the false-positive rates of association have been implemented, wherein FE is applied to loci for which between-study heterogeneity is not detected while RE is applied to loci when between-study heterogeneity is detected[92]. However, Thompson et al.[96] have suggested that FE may not yield p-values that are conservative enough, while RE may estimate overly conservative p-values. Furthermore, Han et al.[100] demonstrated a limitation with the RE model that makes it yield overly conservative p-values. They demonstrated this by a simulation that shows that the FE model is more efficient—that is, $p_{FE} < p_{RE}$—than the RE model at least 75% of the time while the RE model is never more efficient than the FE model because the RE model assumes that τ^2 exists under both the null and the alternative hypotheses with likelihoods L_0 and L_1, respectively

$$L_0 = \prod_i \frac{1}{\sqrt{2\pi(V_i+\widehat{\tau^2})}} \exp\left(-\frac{X_i^2}{2(V_i+\widehat{\tau^2})}\right)$$

$$L_1 = \prod_i \frac{1}{\sqrt{2\pi(V_i+\widehat{\tau^2})}} \exp\left(-\frac{(X_i-\mu)^2}{2(V_i+\widehat{\tau^2})}\right),$$

where $\widehat{\tau^2}$ is the estimated between-study variance and is held constant under L_0 and L_1. In response, Han et al.[100] implemented a modified RE model (RE2 from here forward) that does not assume that heterogeneity estimated under the alternative hypothesis should be applied under the null, providing the likelihoods:

$$L_0 = \prod_i \frac{1}{\sqrt{2\pi(V_i)}} \exp\left(-\frac{X_i^2}{2(V_i)}\right)$$

$$L_1 = \prod_i \frac{1}{\sqrt{2\pi(V_i+\widehat{\tau^2})}} \exp\left(-\frac{(X_i-\mu)^2}{2(V_i+\widehat{\tau^2})}\right).$$

Binary-effects

Han et al.[101] implemented another random-effects based model known as the binary-effects model (BE) that estimates a test statistic for each study in the MA, the m-value. This m-value is equivalent to a Bayesian posterior probability of association, in that it incorporates prior effect existence data from collections. The overall p-value is estimated by incorporating the m-value as a weight in the FE test statistic,

$$Z_{BE} = \frac{\sum_i m_i \sqrt{W_i} z_i}{\sqrt{\sum_i m_i^2 W_i}}.$$

Here m_i, the m-value, is the posterior probability of effect existence in study i,

$$m_i = P(T_i = 1|X) = \frac{P(X|T_i = 1)P(T_i = 1)}{P(X|T_i = 0)P(T_i = 0) + P(X|T_i = 1)P(T_i = 1)}$$

$$= \frac{\sum_{t \in U_i} P(X|T = t)P(T = t)}{\sum_{t \in U} P(X|T = t)P(T = t)}.$$

Here X is the observed effect size, T is a binary random variable that indicates whether an effect exists or not for a given SNP, U is the set of the 2^n values that T can take on, and U_i is the subset of U where T=1. The prior for X, μ, is normally distributed with mean 0 and variance σ^2, where σ was set to 0.2, a value found in simulations[105] to be an appropriate prior that predicts that approximately 5 SNPs per million SNPs will have a high effect size (i.e. OR ~ 3) while the prior for T, π, follows the beta distribution with $\alpha = 1$ and $\beta = 1$ (i.e. uniform distribution). The m-value can be computed analytically by integration or estimated by a Metropolis-Hastings algorithm based Markov Chain Monte Carlo method when the number of studies in a MA does not allow for an analytical solution[101].

Simulations as well as an empirical investigation have shown that the BE model adjusts for moderate ($50\% \geq I^2 \leq 75\%$) between-study heterogeneity slightly better than either RE or RE2[101]. The BE model also provides a visualization framework known as the P-M plot to better understand

the effect existence for each individual study as well as the between-study heterogeneity and evidence of possible allelic heterogeneity (i.e. different minor alleles between subpopulation groups) by plotting the m-value against the p-value.

3.3. Results

For the reanalysis of the Dubois et al. data[72], summary statistics for each of the 12 collections (9,451 cases and 16,434 controls in total) were obtained by fitting a logistic regression model for each collection separately. Study collections had been originally sampled from the UK, Finland, the Netherlands, Italy, USA, Hungary, Ireland, Poland, and Spain. Of the approximately 523,000 SNPs from the discovery stage of the original study, only the 26 SNPs that were genome-wide significant (GWS) in the combined stage of the original study with p-value $\leq 5x10^{-8}$ were included in the present study. For the reanalysis of the Trynka et al. data[76], summary statistics for each of the 7 collections (12,041 cases and 12,228 controls) were obtained by by fitting a logistic regression model that included sex as a covariate for each collection separately. Of the approximately 139,000 SNPs from the dense genotyping study, only the 32 SNPs from the primary signals at a given locus (secondary or tertiary signals within a locus were not included) that were significant with a p-value $\leq 5x10^{-8}$ were included in the present study. To exclude ethnic outliers within each collection and adjust for population stratification, multidimensional scaling (along with other GWAS quality control measures) was performed during data quality control in both of the original studies[106]. All MA under FE, RE, RE2, and BE was performed using the METASOFT package[101].

Replication of Dubois et al. 2010

Between the p-values reported by the FE, RE, RE2, and BE models (Table 3.1), the FE model was the most efficient (i.e. reported the lowest p-value) except when there was moderate to high between-study heterogeneity ($I^2 > 50\%$)[102]. As expected from the simulation results of Han et al.[100],

P_{RE} equaled that of P_{FE} when there was no between-study heterogeneity. Also as expected, the RE model did not yield lower p-values than the FE, RE2, or BE models for any loci. Interestingly, for SNPs rs10903122, rs13010713, rs1464510 and rs653178, P_{RE2} was lower than P_{FE}, P_{RE}, and P_{BE} in the presence of high between-study heterogeneity ($I^2 \geq 75\%$) although the RE2 model still yielded much more conservative p-values than FE-CMH and was not GWS at $p \leq 5x10^{-8}$ for rs10903122 or rs13010713. Finally, as expected based on simulation and empirical evidence reported by Han et al.[101], P_{BE} was lower than P_{RE} for SNPs rs917997 and rs13314993, both of which exhibit moderate between-study heterogeneity ($50\% \geq I^2 \leq 75\%$).

Table 3.1. Meta-analysis p-values of FE, RE, RE2, and BE from Dubois et al.[72]

SNP	Chr:Pos	I^2	P_{Dubois}	P_{FE}	P_{RE}	P_{RE2}	P_{BE}
rs2816316	1:190803436	0	2.20E-17	3.18E-17	3.18E-17	7.79E-17	3.88E-15
rs3748816	1:25116606	4.3	3.28E-09	3.62E-09	1.40E-08	6.51E-09	2.61E-08
rs10903122	1:25176163	70.3	1.73E-10	7.84E-04	8.57E-02	6.59E-06	3.08E-05
rs296547	1:199158760	2.3	4.11E-09	5.11E-09	9.76E-09	9.12E-09	3.33E-08
rs13003464	2:61040333	16.4	3.71E-13	4.24E-13	6.02E-11	9.08E-13	1.14E-11
rs917997	2:102437000	64.2	1.11E-15	1.07E-15	5.57E-05	1.56E-16	1.99E-17
rs13010713	2:181704290	64.0	4.74E-11	1.58E-05	2.90E-02	6.72E-07	7.26E-07
rs4675374	2:204510823	0	5.79E-09	6.41E-09	6.41E-09	1.14E-08	1.09E-07
rs17035378	2:68452459	0	7.79E-09	8.11E-09	8.11E-09	1.51E-08	2.55E-07
rs13098911	3:46210205	3.4	3.26E-17	4.07E-17	3.61E-16	9.44E-17	6.37E-16
rs17810546	3:161147744	17.4	3.98E-28	6.78E-28	3.46E-22	2.10E-27	1.74E-25
rs1464510	3:189595248	92.1	2.98E-40	6.26E-12	1.76E-01	4.38E-33	1.74E-29
rs13314993	3:32990473	61.5	3.27E-09	2.94E-05	6.45E-02	1.70E-06	1.06E-06
rs11712165	3:120601486	35.1	8.03E-09	1.24E-08	4.44E-06	2.25E-08	9.62E-08
rs13151961	4:123334952	0	2.18E-27	3.08E-27	3.08E-27	9.44E-27	2.26E-24
rs2327832	6:138014761	0	4.46E-19	5.86E-19	5.86E-19	1.51E-18	1.63E-16
rs1738074	6:159385965	0	2.94E-15	3.63E-15	3.63E-15	8.35E-15	8.56E-13
rs10806425	6:90983333	27.2	3.89E-10	4.93E-10	1.69E-07	9.21E-10	9.70E-09
rs802734	6:128320491	5.8	2.62E-14	3.65E-14	2.18E-13	8.11E-14	2.81E-12
rs653178	12:110492139	87.1	7.15E-21	1.58E-03	5.72E-01	2.92E-14	1.10E-10
rs9792269	8:129333771	0	3.28E-09	2.98E-09	2.98E-09	5.37E-09	8.53E-08
rs1250552	10:80728033	21.0	9.09E-10	5.45E-08	3.88E-06	9.87E-08	1.31E-07
rs1893217	18:12799340	0	2.52E-10	2.68E-10	2.68E-10	5.11E-10	2.49E-08
rs11221332	11:127886184	0	5.28E-16	7.74E-16	7.74E-16	1.92E-15	7.35E-14
rs12928822	16:11311394	14.5	3.12E-08	2.91E-08	9.00E-07	5.68E-08	2.52E-07
rs4819388	21:44471849	39.7	2.46E-09	2.92E-09	2.39E-06	3.40E-09	5.20E-09

Representative P-M plots were generated for three loci GWS in Dubois et al.[72] that represented either zero, moderate, or high degree of between-study heterogeneity: rs13151961 (I^2 = 0), rs917997 (I^2 = 64.2), and rs1464510 (I^2 = 92.1). The P-M plot for rs13151961 (Figure 3.1) reflects the lack of between-study heterogeneity in effect size by showing consistently high m-values against consistently non-GWS (for each individual collection) p-values across FE-CMH, FE, RE, RE2, and BE. For SNP rs917997 (Figure 3.2), the m-values were spread more widely across the horizontal axis of the P-M plot, with the Spanish, Irish, and Italian collections occupying what Han et al.[101] refer to as the ambiguous region of effect existence. The impact of the between-study heterogeneity and the clustering of collections in the ambiguous effect existence region may help explain the much larger p-values for RE, with BE yielding the lowest p-value. SNP rs1464510 (Figure 3.3) exhibited the greatest degree of between-study heterogeneity with collections clustering to the relatively unambiguous regions of no effect existence or high-probability of effect existence with the exception of two studies (m-value > 0.8) that lie in the ambiguous region. The combined stage p-value reported by Dubois et al.[72] for rs1464510 was GWS, as were FE, RE2, and BE (with RE2 yielding a lower p-value than FE or BE), while RE was not surprisingly non-GWS.

In the discovery stage of Dubois and colleagues' association analysis, rs10903122, a SNP of the *RUNX3* gene, was one of the SNPs to pass the discovery stage p-value threshold of less than $1x10^{-4}$ under FE-CMH and was one of the 131 SNPs that were subsequently genotyped in the replication and the final combined GWAS. SNP rs10903122 was then reported as being GWS with a combined p-value < $5x10^{-8}$. However, when between-study heterogeneity was adjusted for in a replication of the discovery stage of Dubois et al.[72], rs10903122 (I^2 = 70.3) did not pass the p-value threshold of $1x10^{-4}$ set by Dubois et al.[72] under RE, RE2, or BE with the lowest p-value under RE2 being $3.91x10^{-3}$. Finally, in the replication of the combined stage, rs10903122 did not meet the combined stage GWS threshold of $5x10^{-8}$, with the lowest p-value obtained under RE2 being

$6.59x10^{-6}$. While the p-values indicate non-GWS, the P-M plot of rs10903122 (Figure 3.4) provides evidence that there may be a true effect, as the m-values for six of the twelve collections are within the region of high posterior probability of effect existence (m-value ~ 0.9) while another four collections are in an area of the P-M plot where there is overlap between the ambiguous region of effect existence and high posterior probability of effect existence[101]. Only two collections, the Polish and Dutch collections, have an m-value indicative of no effect existence. From the 26 SNPs that were GWS in the combined stage of Dubois et al.[72] with a p-value $\leq 5x10^{-8}$, five SNPs were not GWS under the FE, RE, RE2, or BE model. As only one of the 26 SNPs had a lower p-value using FE, RE, RE2, or BE, the least conservative method is FE-CMH. As expected, the RE model was the most conservative in the presence of between-study heterogeneity.

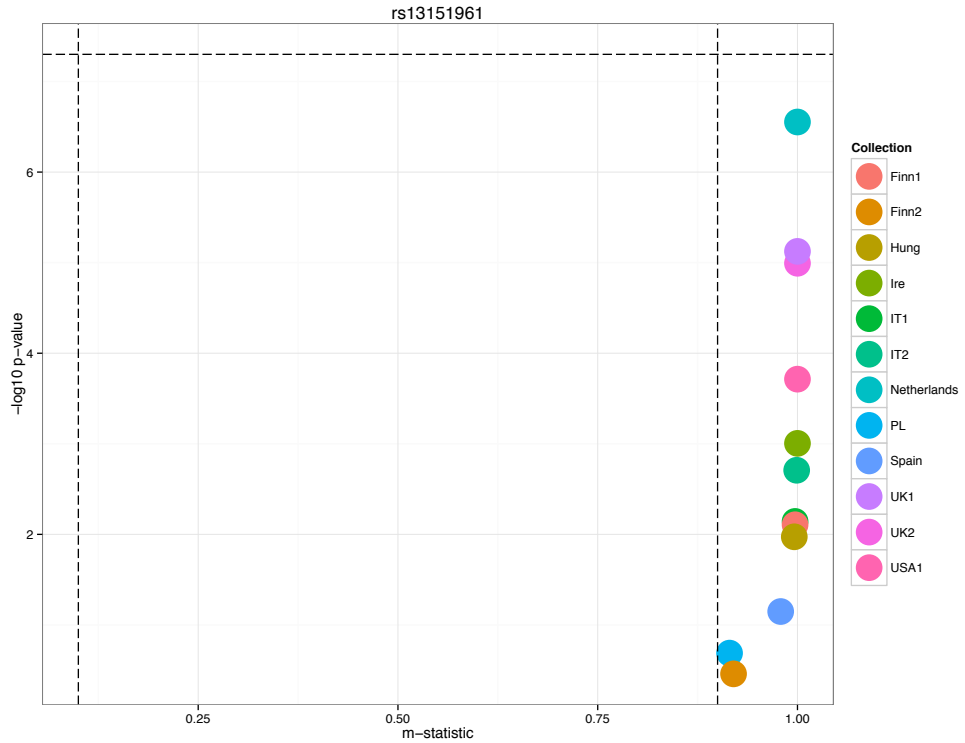


Figure 3.1. P-M plot of rs13151961 from Dubois et al.[72]

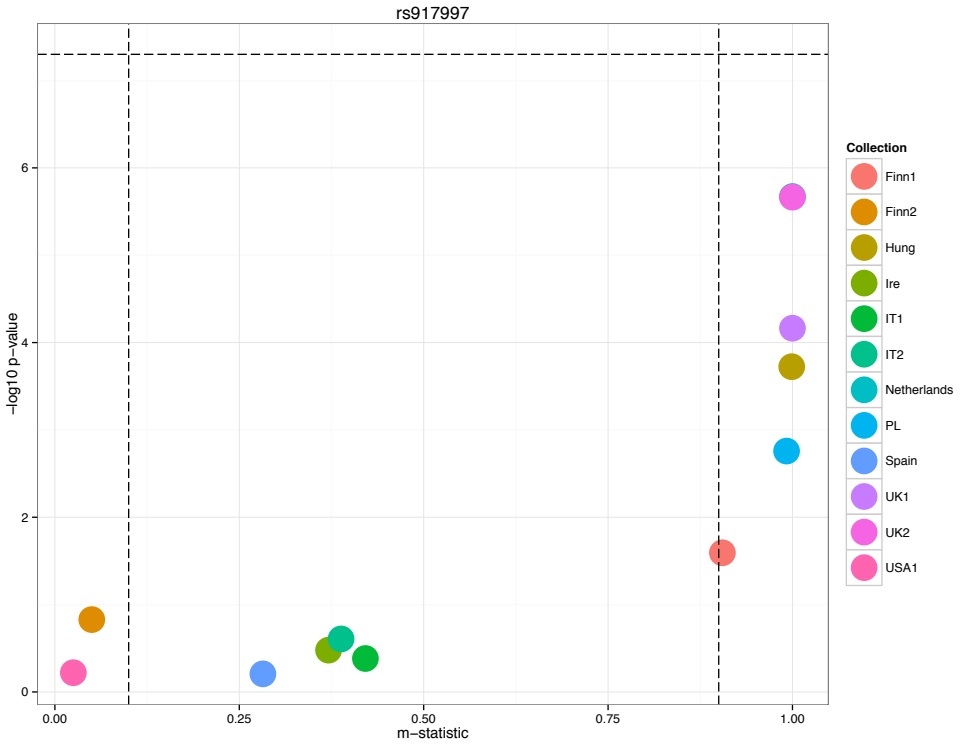


Figure 3.2. P-M plot of rs917997 from Dubois et al.[72]

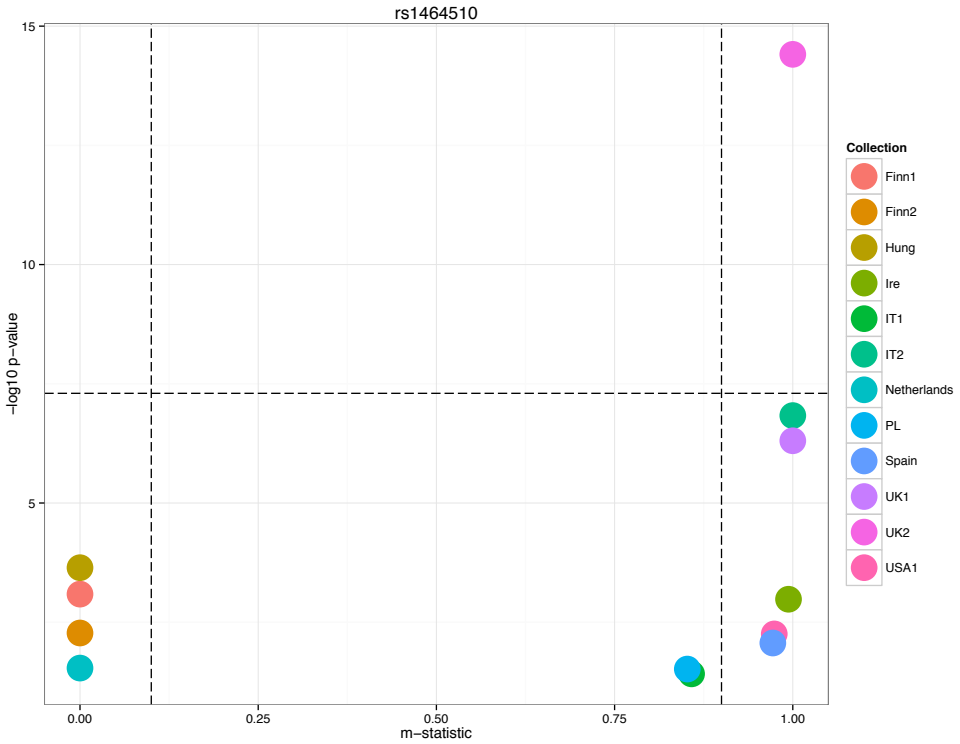


Figure 3.3. P-M plot of rs1464510 from Dubois et al.[72]

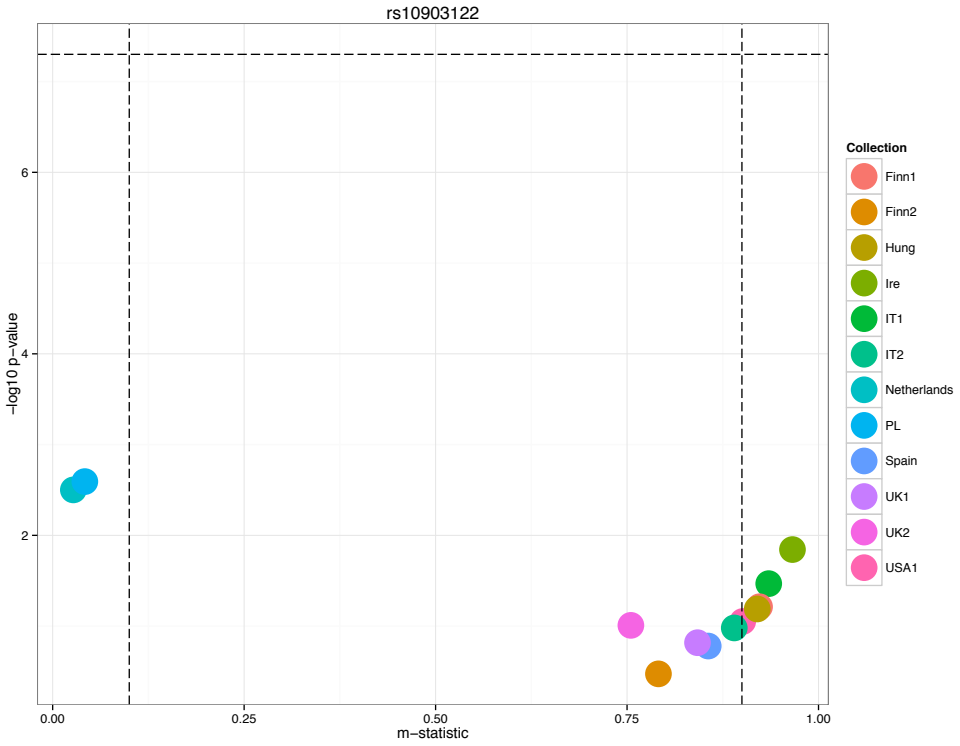


Figure 3.4. P-M plot of rs10903122 from Dubois et al.[72]

Replication of Trynka et al. 2011

Results for the 32 SNPs tested from the reanalysis of the Trynka et al.[76] data are shown in Table 3.2. Not surprisingly, the least conservative method was FE-CLR and only 3 of the 32 tested SNPs exhibited a lower p-value under FE, RE, RE2, or BE—however, the differences in p-values were marginal. Between FE, RE, RE2, and BE, FE was the most efficient model except in the presence of moderate to high heterogeneity ($I^2 > 50\%$). As was demonstrated in the simulation results of Han et al.[101] and the reanalysis of Dubois et al.[72], performance of RE matches that of FE when no between-study heterogeneity is detected and RE never outperformed FE, RE2, or BE for any loci. For the highly heterogeneous SNP, rs1050976, RE2 outperformed FE, RE, and BE by several orders of magnitude and was only marginally less efficient than FE-CLR and just missed the GWS threshold of p-value $\leq 5 \times 10^{-8}$. There was one anomalous SNP, rs3184504, that showed unexpected

behavior as BE outperformed RE2 in the presence of high between-study heterogeneity ($I^2 > 75\%$). This was considered anomalous because Han et al.[101] has previously demonstrated through simulation that while BE and RE2 are nearly the same in efficiency, RE2 tends to perform slightly better than BE in the presence of high between-study heterogeneity.

Table 3.2. Meta-analysis of FE, RE, RE2, and BE from Trynka et al.[76]

SNP	Chr:Pos	I^2	P_{Trynka}	P_{FE}	P_{RE}	P_{RE2}	P_{BE}
rs4445406	1:2539400	0	5.40E-12	1.98E-11	1.98E-11	4.40E-11	1.57E-10
rs12068671	1:172681031	41.7	1.40E-10	4.85E-10	7.28E-03	4.55E-10	9.45E-11
rs1359062	1:192541472	0	2.50E-25	1.57E-23	1.57E-23	4.99E-23	1.25E-22
rs10800746	1:200881392	0	2.60E-08	4.00E-08	4.00E-08	7.15E-08	2.59E-07
rs13003464	2:61186829	0	4.30E-16	7.83E-16	7.83E-16	2.05E-15	2.21E-14
rs990171	2:103086770	0	1.20E-16	1.79E-15	1.79E-15	4.62E-15	1.91E-14
rs1018326	2:182007800	60.4	3.10E-16	3.48E-15	4.07E-02	1.78E-15	2.12E-16
rs6715106	2:191913034	58.3	8.40E-09	4.44E-08	1.32E-03	7.86E-08	2.00E-07
rs1980422	2:204610396	34.2	1.40E-15	3.02E-15	4.55E-07	7.74E-15	2.46E-14
rs2097282	3:46378025	54	1.10E-20	2.46E-19	1.45E-05	7.10E-19	4.23E-19
rs61579022	3:119123278	49.1	9.90E-09	1.75E-08	1.99E-02	3.16E-08	1.06E-08
imm_3_161120372	3:157034177	0	2.60E-27	1.97E-26	1.97E-26	6.61E-26	5.96E-25
rs2030519	3:188119901	0	3.00E-49	2.36E-48	2.36E-48	1.06E-47	1.03E-46
rs13132308	4:123551114	0	1.90E-38	6.99E-36	6.99E-36	2.72E-35	1.82E-34
rs1050976	6:408079	83.1	1.80E-09	3.47E-03	3.77E-01	8.08E-08	2.42E-05
rs55743914	6:128293562	29.6	1.10E-18	9.78E-17	3.53E-06	2.63E-16	9.34E-17
rs17264332	6:138005515	0	5.00E-30	9.83E-28	9.83E-28	3.38E-27	3.17E-26
rs182429	6:159469574	0	8.50E-16	1.16E-15	1.16E-15	3.01E-15	3.06E-14
1kg_7_37384979	7:37298049	63.2	2.10E-08	8.87E-09	9.67E-02	9.87E-09	1.94E-09
rs2387397	10:6390192	0	1.90E-08	4.42E-08	4.42E-08	7.82E-08	7.19E-07
rs1250552	10:81058027	59	8.00E-17	1.56E-14	8.57E-04	3.91E-14	1.82E-14
rs7104791	11:111196858	0	1.90E-11	2.70E-10	2.70E-10	5.71E-10	6.29E-10
rs10892258	11:118579865	0	1.70E-11	5.86E-11	5.86E-11	1.28E-10	3.17E-10
rs61907765	11:128391937	0	3.40E-13	4.47E-13	4.47E-13	1.06E-12	1.43E-11
rs3184504	12:111884608	81.6	5.40E-21	2.20E-13	4.25E-01	1.98E-16	3.80E-17
rs11851414	14:69259502	0	4.70E-08	7.97E-08	7.97E-08	1.44E-07	3.46E-07
rs1378938	15:75096443	0	7.80E-09	3.26E-07	3.26E-07	5.68E-07	2.06E-06
rs6498114	16:10964118	0	5.80E-10	3.17E-10	3.17E-10	6.69E-10	1.76E-09
rs11875687	18:12843137	8.17	1.90E-10	3.51E-10	1.33E-07	7.39E-10	3.03E-09
rs1893592	21:43855067	15.5	3.00E-09	7.07E-09	4.30E-04	1.26E-08	4.99E-09
rs4821124	22:21979289	0	5.70E-11	8.58E-11	8.58E-11	1.86E-10	6.36E-10
rs13397	X:153248248	0	2.70E-08	7.33E-09	7.33E-09	1.44E-08	1.58E-08

P-M plots were generated for SNPs with the highest between-study heterogeneity and the lowest between-study heterogeneity: rs1050976 (I^2 = 83.1), rs3184504 (I^2 = 81.6,), and rs2030519 (I^2 = 0). For SNP rs1050976 (Figure 3.5), all collections, with the exception of the UK collection, cluster towards the bottom left-hand side of the plot where the posterior probability of effect existence is essentially zero. It is of note, that the UK collection, with an m-value of approximately 1.0 and near GWS p-value, is at the top right-hand side of the plot where the posterior probability of effect existence approaches 100%. Such a divergence suggests that much of the between-study heterogeneity observed at rs1050976 is being driven by the considerable divergence between the UK collection and the six other collections. While the clustering to either end of the m-value distribution is not quite as distinct as with rs1050976, the P-M plot for rs3184504 (Figure 3.5) suggests that the high between-study heterogeneity of rs3184504 is probably driven by the UK collection again having a very high posterior probability of effect existence and the clustering of the Spanish and Dutch collections towards the region of effect existence. The standout collection here was the Indian collection, which is very near the bottom-center of the P-M plot and indicates that the Indian collection is underpowered[101]. This suggestive evidence from the P-M plot is corroborated by the very small sample size of the Indian collection (229 cases and 391 controls) relative to the total sample size of the study (12,041 cases and 12,228 controls). SNP rs2030519 (Figure 3.6) shows a distinct clustering of all points towards the region of effect existence even though the individual study p-values are not GWS for most of the collections, again with the exception of the UK collection that alone has a p-value < $1x10^{-30}$. It is this distinct clustering of all collections to the region of effect existence that probably drove the observed zero between-study heterogeneity.

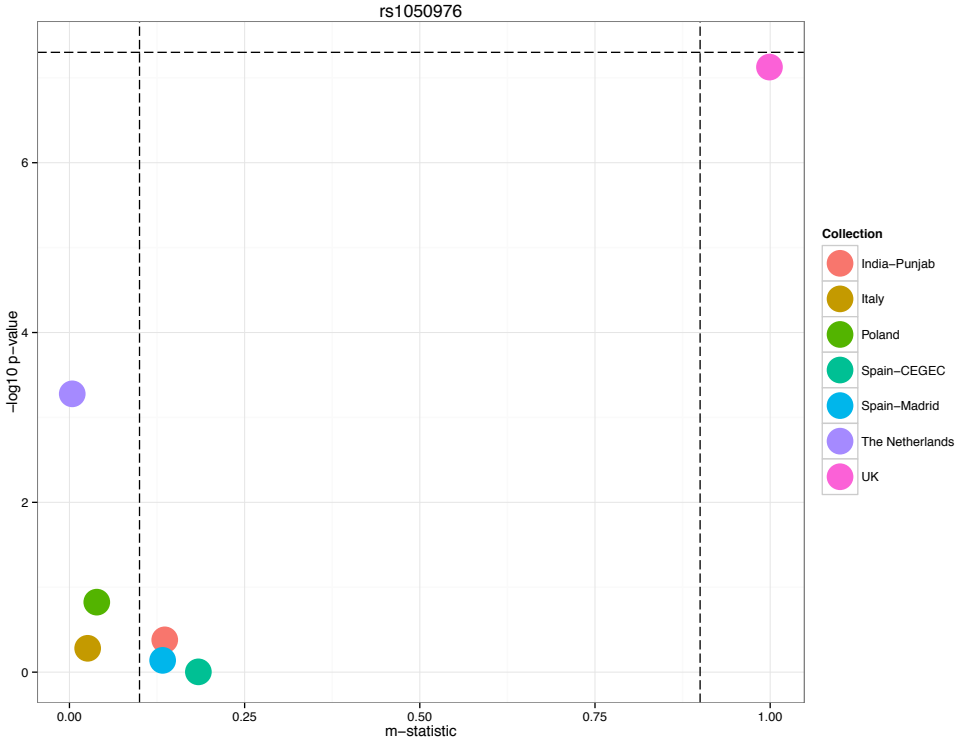


Figure 3.5. P-M plot of rs1050976 from Trynka et al.[76]

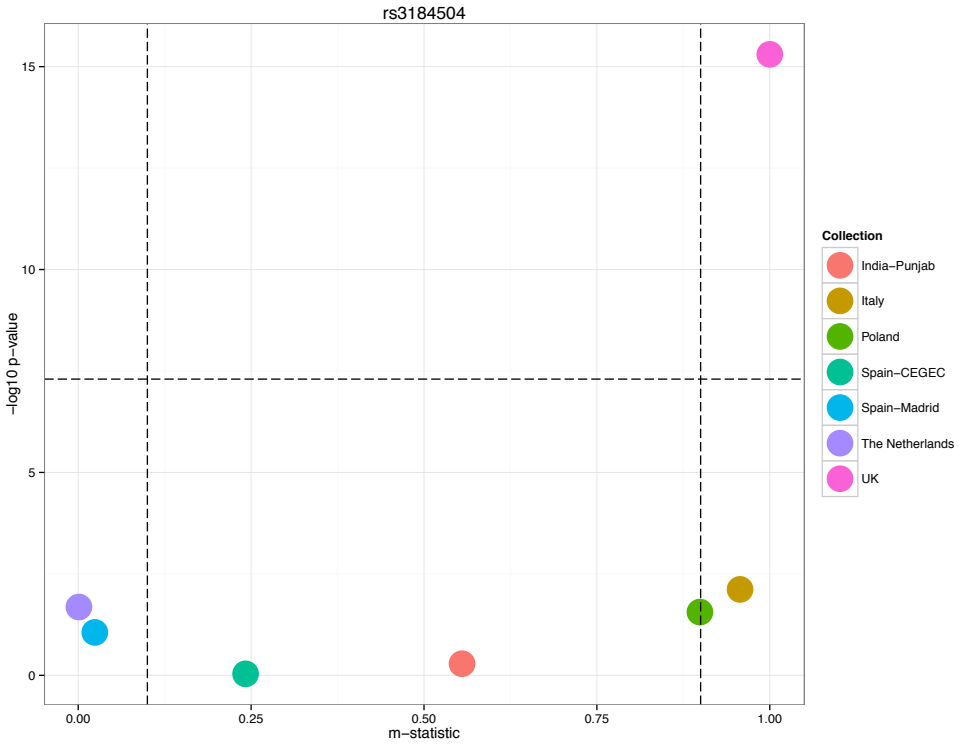


Figure 3.6. P-M plot of rs3184504 from Trynka et al.[76]

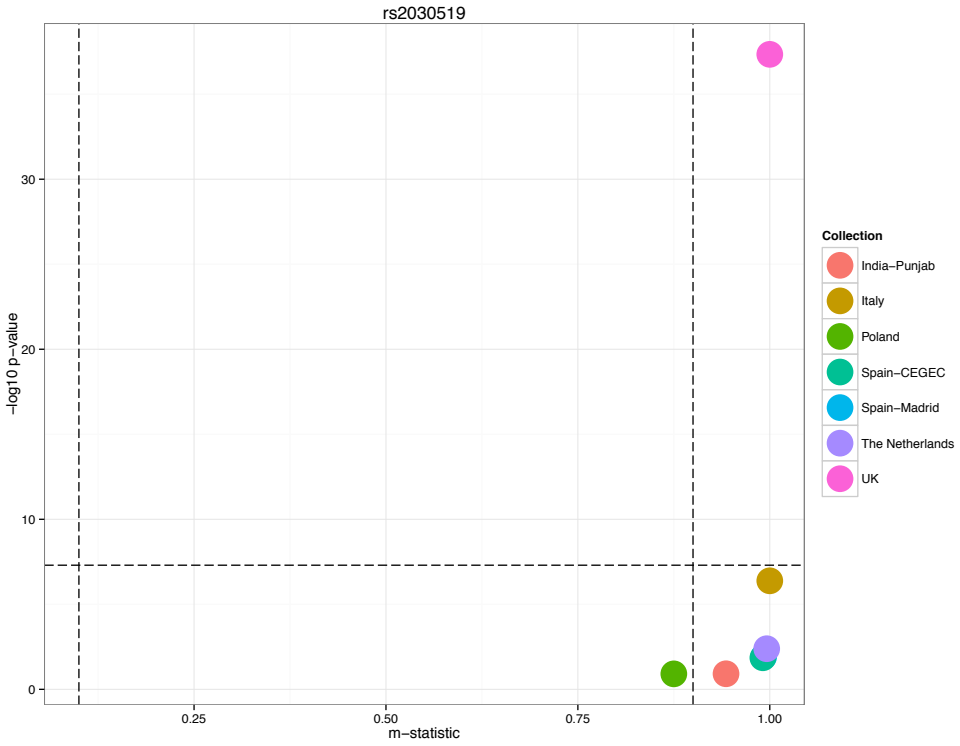


Figure 3.7. P-M plot of rs2030519 from Trynka et al.[76]

To investigate allelic heterogeneity, the impact of effect direction and the minor allele at a given locus were considered (Table 3.3). For the least heterogeneous SNP, rs2030519, effect direction and the minor allele are homogeneous (OR < 1, minor allele = G) across all collections while the two SNPs with the highest degrees of between-study heterogeneity have mixed effect direction and minor alleles. The minor alleles for both rs1050976 and rs3184504 differ for the collections with m-values approaching 1.0. To further investigate the high between-study heterogeneity of rs1050976 and rs3184504, wherein the between-study heterogeneity is likely to be driven by the effect existence divergence of the UK collection alone, MA on the Trynka et al. data was performed without the UK collection (the m-value is nearly 1 across all loci for the UK collection.). Of the 32 SNPs that were originally GWS under FE-CLR, only three SNPs are GWS when the UK collection is excluded from the MA, including imm_3_161120372 ($P_{FE} = 2.18 \times 10^{-9}$,

$P_{RE} = 2.18x10^{-9}$, $P_{RE2} = 3.62x10^{-9}$, $P_{BE} = 2.87x10^{-9}$), rs2030519 ($P_{FE} = 2.48x10^{-12}$, $P_{RE} = 2.48x10^{-12}$, $P_{RE2} = 4.72x10^{-12}$, $P_{BE} = 5.85x10^{-11}$) and rs13132308 ($P_{FE} = 3.99x10^{-8}$, $P_{RE} = 3.99x10^{-8}$, $P_{RE2} = 6.66x10^{-8}$, $P_{BE} = 6.90x10^{-7}$). Interestingly, all three SNPs that were found to be GWS when the UK collection was excluded also had zero between-study heterogeneity, providing suggestive evidence that the UK collection was driving up the between-study heterogeneity.

Table 3.3. Minor alleles, allele frequency, and odds ratio with 95% confidence intervals for three representative SNPs.

SNP	Collection	Minor Allele	Minor Allele Frequency*	Odds Ratio	95% Confidence Interval
rs1050976	India-Punjab	T	0.40	0.89	0.68 – 1.17
	UK	C	0.47	1.13	1.08 – 1.18
	Spain-Madrid	T	0.48	0.97	0.79 – 1.18
	Poland	T	0.44	0.88	0.74 – 1.05
	Spain-CEGEC	T	0.47	1.00	0.82 – 1.23
	The Netherlands	T	0.47	0.77	0.66 – 0.89
	Italy	T	0.46	0.96	0.86 – 1.08
rs3184504	India-Punjab	T	0.13	0.88	0.61 – 1.28
	UK	C	0.51	0.83	0.79 – 0.87
	Spain-Madrid	T	0.47	1.19	0.97 – 1.46
	Poland	C	0.50	0.82	0.68 – 0.98
	Spain-CEGEC	T	0.49	0.99	0.81 – 1.20
	The Netherlands	T	0.47	1.19	1.03 – 1.39
	Italy	C	0.49	0.86	0.77 – 0.96
rs2030519	India-Punjab	G	0.40	0.81	0.62 – 1.06
	UK	G	0.48	0.74	0.71 – 0.78
	Spain-Madrid	G	0.53	0.77	0.63 – 0.94
	Poland	G	0.46	0.87	0.72 – 1.04
	Spain-CEGEC	G	0.53	0.78	0.64 – 0.95
	The Netherlands	G	0.46	0.80	0.69 – 0.93
	Italy	G	0.53	0.75	0.67 – 0.84

*Minor allele frequency in controls

3.4. Discussion

In two previous GWAS MA of celiac disease, Dubois et al.[72] and Trynka et al.[76] presented evidence that *RUNX3* is a novel locus associated with celiac disease. However, when between-study heterogeneity was accounted for in the present study by MA models such as RE, RE2, and BE, the SNPs identified as GWS in the *RUNX3* gene no longer provide convincing evidence in terms of naïve p-value alone. As this locus was only marginally significant under all FE, RE, RE2, and BE models (and not GWS under even RE2 as described above) and exhibited the highest degree of between-study heterogeneity (I^2 = 83.1), this particular locus may require an additional validation to be deemed a new independent loci associated with celiac disease strictly by p-value alone. However, the m-values and the P-M plot for rs10903122 provide suggestive evidence of a true effect and that the sample sizes for most of the collections are large enough, even if the posterior probability of effect existence for those collections are close to zero. Interestingly, three of the GWS novel loci (rs12068671, 1kg_7_37384979, and rs1893592) from Trynka et al.[76] that were analyzed under FE-CLR were also GWS under FE, RE2, and BE in the present study. Four loci for which the BE model was most efficient (rs1018326, rs61579022, rs55743914, and rs31854504), demonstrate evidence that even with between-study heterogeneity observed, those loci do replicate the results shown previously in Dubois et al.[72]. For the loci tagged by rs10903122 and rs13314993, which are not GWS at p-value $\leq 5x10^{-8}$ under FE, RE, RE2, or BE in the replication of the combined stage of the Dubois study, but are GWS under FE-CMH, those same loci do not have any SNPs GWS under FE-CLR in Trynka et al.[76]. However, SNPs representing two loci, rs653178 and rs917997, were found to be GWS under FE-CLR in Trynka et al.[76]. Curiously, the most heterogeneous SNP from the Dubois et al.[72] data, rs1464510, has zero observed between-study heterogeneity with rs2030519 in the Trynka et al.[76] data. Both SNPs belong to the *LPP* locus and both SNPs are GWS under FE-CMH and FE-CLR as well as FE, RE2, and BE.

Trynka et al.[76] reported that the UK collection alone was sufficient to account for most of the GWS loci but did not present data in their study. The present study provides evidence that this assertion is true and that the UK collection is driving much of the evidence for effect existence and the high degree of between-study heterogeneity observed in some loci and that the MA based estimates of the effect sizes are probably biased upwards because of the "winner's curse" phenomenon[35,107–109]. There are two loci, rs1893592 and rs12068671, for which logistic regression using just the UK collection alone slightly outperforms FE-CLR, using all 7 collections, although with rs12068671, $P_{BE} < P_{\text{FE-CLR UK only}}$ probably because of the moderate amount of between-study heterogeneity that the BE model can adjust for.

Strengths of the original meta-analyses by Dubois et al.[72] and Trynka et al.[76] include homogeneity of the phenotyping of samples and of the genotyping platforms for each study dataset, especially Trynka et al.[76] dataset, which was based on one genotyping platform. While the collections for the Trynka dataset were genotyped in different laboratories and may still be subject to some laboratory-level bias, a bias due to genotyping platform[93] is not likely to exist because all samples were genotyped on the same Illumina Immunochip platform. However, this study demonstrates that even without the genotyping platform bias, significant between-study heterogeneity exists and does affect the performance of the MA models. As population stratification was adjusted for by multi-dimensional scaling, the observed heterogeneity is likely to be driven by ethnic or subpopulation group membership.

The present study lacks independent samples to further validate the results, particularly any of the novel signals from the Trynka et al.[76] study. This weakness is present in many MAs of GWAS as avoiding the genotyping of more samples is the impetus for MA in the first place. For instance, with rs1050976 and rs3184504, there were no additional independent samples to investigate whether small sample size is driving the apparently high between-study heterogeneity. While these SNPs

could be forward imputed into other large GWAS datasets, it is not clear that the between-study heterogeneity will be reduced, especially if imputation quality is low. Also, this present study did not investigate the secondary and tertiary signals at loci identified by fine-mapping in the Trynka et al.[76] study and we did not perform MA by FE, RE, RE2, and BE on all SNPs from each dataset. Finally, the I^2 statistic is based on Cochran's Q statistic and if $Q < (k - 1)$, where k is the number of studies, then the heterogeneity truncates to zero[103]. Although the data suggests that heterogeneity below 50% has little effect on the overall p-value, the possibility of a false negative cannot be ruled out.

The present study is the first study that the author is aware of that applies FE and RE based MA models for GWAS MA of celiac disease, including the newly developed BE model and the P-M plot framework developed by Han et al.[101]. This case study demonstrates that if Dubois et al.[72] and Trynka et al.[76] performed their respective studies with either FE, RE, RE2, or BE, rs1050976, rs10903122 and rs13314993 from the *RUNX3* gene would not have been identified as GWS at $p < 5x10^{-8}$, most likely because of the high degree of between-study heterogeneity that exists at those sites, although they may have been presented as loci with suggestive evidence of association. However, and perhaps most importantly, this study demonstrates that FE-CMH and FE-CLR with ethnic collection membership as a covariate is effectively equivalent to the ideal FE and RE hybrid strategy that outlined by Han et al.[101].

Chapter 4: Association Analysis of the Extended MHC Region in Celiac Disease

The very strong genetic effects on celiac disease from the known HLA high-risk loci, and the complex nature of the major histocompatibility complex (MHC), have greatly complicated any thorough statistical genetic analysis of the region. The purpose of this study was to test the hypothesis that additional novel celiac disease loci exist within the extended MHC (xMHC). A total of 1898 SNPs were tested for association with celiac disease across the 7.6 Mb xMHC region using 1668 cases and 517 controls. A conditional inference based recursive partitioning method was implemented to create an informative factor variable of the known *HLA-DQA1* and *HLA-DQB1* high-risk genotypes that was included in a multiple logistic regression model for association testing. A linkage disequilibrium (LD) based fine-mapping method was implemented to estimate the number of independent celiac disease loci present in the xMHC after accounting for the known HLA effects. Four novel and independent celiac disease loci were found to be statistically significant within the classic MHC region. This is the first comprehensive celiac disease association analysis of the xMHC that accounts for the known HLA disease genotypes and the genetic complexity of the region.

4.1. Introduction

Celiac disease is a common T cell mediated, auto-immune disorder that is triggered by ingestion of dietary gluten. The population prevalence of the disease (which occurs primarily amongst Caucasians) is approximately 1%[110], with mounting evidence in the literature that suggests that the incidence of disease is increasing[111,112]. Comorbid diseases that occur alongside celiac disease include autoimmune disorders such as type I diabetes, autoimmune thyroiditis, inflammatory bowel disease, and adult rheumatoid arthritis[113–116].

Association between histocompatibility antigens in the major histocompatibility complex (MHC) and celiac disease were first documented over 40 years ago[117,118] followed by the identification of the HLA-DQ2 molecule about a decade later[119]. While several linkage studies were performed to identify highly penetrant genes, no other consistent high-risk loci were found other than at the HLA loci[67,120–126]. Within the last decade, genome-wide association studies (GWAS) and follow-up studies of celiac disease have identified and replicated 39 non-HLA loci that are associated with celiac disease, explaining about 5% of the estimated disease risk[68,69,71–73]. The strongest GWAS associations were in the MHC region, with the most strongly associated SNP explaining about 35% of the disease risk.

HLA class II molecules such as the DQ2 molecule are necessary components in the development of celiac disease by encoding the cell surface proteins on CD4+ T lymphocytes that recognize gliadin, a component protein of gluten[127]. The specific DQ serotype expressed is determined by the alleles in the HLA class II genes, *HLA-DQA1* and *HLA-DQB1*. Over 90% of celiac disease cases express HLA DQ2[58,128–130]. Another 5% of celiac disease cases express DQ8[131–133], while the remaining 3–5% of celiac disease cases carry neither DQ2 or DQ8, although most of these cases will carry at least the DQB1*02 allele. However, as approximately 30% of Caucasians carry the genotype to express HLA DQ2, but only about 1% develop the disease, the HLA association is considered to be necessary but not sufficient in the etiology of celiac disease[134–136].

This study represents the first investigation of the extended MHC (xMHC) region of chromosome 6 for additional, non-HLA, disease-associated common variants. Studies of the non-HLA associations in the MHC have been conducted in related autoimmune diseases such as systemic lupus erythematosus (SLE)[137] and type 1 diabetes (IDDM)[138–140]. These studies have found evidence for novel, HLA independent genetic associations within the MHC region. Although the specific statistical methods implemented in these previous studies are study-specific, these studies

shared a common methodological approach, wherein the underlying information about the known high-risk HLA alleles was captured to account for the known risk alleles in an association analysis. In the present study, a similar approach was taken to account for the known HLA risk alleles and test the hypothesis that there are additional disease-associated common variants in the xMHC region other than the known *HLA-DQA1* and *HLA-DQB1* disease alleles. This chapter was adapted from a previously published manuscript in the journal, PLOS ONE[74].

4.2. Methods

Conditional inference based recursive partitioning

A conditional inference based recursive partitioning method was used to partition individuals into strata based on combinations of their *HLA-DQA1* and *HLA-DQB1* genotypes that minimized within-strata heterogeneity. Recursive partitioning is a two-stage process in which predictor variables (in this case, the HLA high-risk genotypes) are selected and then the sample is subjected to a series of binary splits[141]. In this study, to select the predictor variables to be included in the recursive partitioning model, a global test of independence between all the input variables and the outcome was carried out. If the null hypothesis of independence could be rejected at a pre-determined p-value threshold of 0.05, the input variable with the strongest association to the response was selected. The extent of association was measured by the p-value corresponding to a test of the partial null hypothesis of a single input variable and the response variable. A binary split was then imposed on the selected input variable. These steps were repeated until the global null hypothesis of independence could be rejected. A five-level factor variable was created from the terminal nodes of the binary inference tree resulting from the binary splitting process. The terminal nodes of the inference tree determined which strata each individual belonged to with respect to their *HLA-*

DQA1 and *HLA-DQB1* genotypes. The conditional inference based recursive partitioning was performed using the PARTY package in R[142].

Linkage disequilibrium (LD)-based SNP Grouping

A data-reduction method was implemented to group sets of SNPs that had highly correlated associations with celiac disease and determine how many of the SNPs showing significant association were likely to be independent with respect to LD, i.e., in linkage equilibrium. The SNPs were grouped together into LD groups (or 'clumps') by p-values and LD (measured by r^2) in sliding windows, wherein the SNP with the most statistically significant association, or 'index' SNP, was considered an independent locus. The non-index SNPs had to meet a pre-specified secondary p-value threshold and physical distance threshold (in kilobases from the index SNP) in addition to the LD threshold for each LD-based group. The 'clump' procedure in PLINK[143] was used to compute the LD-based SNP grouping analysis.

Association analysis

Logistic regression models were used to perform the association analysis. The genotypes 1/1, 1/2 and 2/2 were unphased and encoded as 0, 1 or 2 to indicate the number of minor alleles present. The multi-level factor variable computed to account for the *HLA-DQA1* and *HLA-DQB1* haplotypes was included in a multiple regression model along with the SNP genotype. All simple and multivariate regression models were fit to the data using PLINK and the GenABEL package in R[143,144].

Study subjects

As part of the celiac disease GWAS, the North American Celiac Disease Genetic consortium was formed and is comprised of Dr. S. Neuhausen at COH and Dr. C. Garner at UC Irvine, Dr. J. Murray at the Mayo Clinic, Dr. A. Fasano at the University of Maryland, and Dr. P. Green at Columbia University with study subjects enrolled in previous studies at each center. Subjects had

already been coded and made non-identifiable. The City of Hope IRB approved this study in January 2010 (protocol 09169). All Mayo Clinic subjects had been previously enrolled under an approved IRB protocol 1173-99. Samples from the University of Maryland were collected under three approved IRB protocols (H-27784, H29090, and H-29938). Samples from Columbia University were collected under an approved IRB protocol AAAE8893 (as well as a previous protocol 8562). Written informed consent was collected from each participant as was described in all of the protocols.

All of the 2300 subjects were Caucasian as nearly all celiac disease cases are of Caucasian descent. Of the cases, 532 were from COH, 743 from the Mayo Clinic, 423 from the University of Maryland, and 66 from Columbia University for a total of 1764 cases. Of the controls, 177 were from COH and 359 from the Mayo Clinic for a total of 536 controls. Blood samples were collected for serological testing and extraction of DNA for genetic studies. All sites used the same serological test, a similar questionnaire, and the same criteria for diagnosing celiac disease. The following criteria had to be met to be defined as a celiac disease case: 1) test positive for a celiac disease specific autoantibody (IgA EMA and IgA tTG antibodies); 2) a proximal small intestinal biopsy that is compatible with celiac disease; 3) clinical and/or histological improvement under a gluten-free diet. While the majority of cases fulfilled all three criteria, a small proportion of subjects that were diagnosed before modern serology came into use, did not have a celiac disease specific serology, while another small minority of subjects did not have a small intestinal biopsy. As sensitivity and specificity for the IgA tTG and IgA EMA tests was reported be nearly 100%, those who tested positive for both IgA tTG and IgA EMA were considered to be positive for celiac disease[145,146]. A small intestinal biopsy was performed on about 90% of those who had a positive serology, and all biopsies were positive for celiac disease. All self-reported celiac cases were not included as cases. All unaffected controls had a negative serology for celiac disease.

HLA typing

The first step in determining the *HLA-DQA1* and *HLA-DQB1* alleles in the cases and controls was to directly sequence the HLA genotypes of 95 individuals. These 95 samples were positive controls for the high-throughput HLA genotyping of the remaining case and control individuals. Direct sequencing of the second exons of *HLA-DQA1* and *HLA-DQB1* by Sanger sequencing was performed using the ABI Prism BigDye Terminator cycle sequencing kit 3.1 (PE Applied Biosystems). Sequence alignment was performed using the Sequencher software (GeneCode Corporation, MI) and manually inspected as necessary. Sample DQ allele assignments were manually compared to DQ alleles from the dbMHC database (http://www.ncbi.nlm.nih. gov/gv/mhc). Using the samples with DQ alleles determined by direct sequencing, *HLA-DQA1* and *HLA-DQB1* genotypes for the remaining samples were determined by one of two high-throughput DQ typing methods. 389 samples (from COH) were genotyped by an allele-specific PCR method developed in the Neuhausen laboratory by Feolo et al.[147]. The genotyping accuracy for this PCR method was greater than 98%. A highly sensitive and specific tag SNP approach was implemented to genotype the remaining 1816 samples by using six SNPs to predict the four *HLA-DQ* types (DQ2.5, DQ2.2, DQ7, and DQ8) that are known to be associated with celiac disease[148]. Genotyping call rates ranged between 95% and 99% while duplicate concordance rates were greater than 99%. The concordance rate between direct sequencing and the tag SNP approach for the 95 samples that were directly sequenced was 100%.

Genotype data

2300 samples plus duplicates were genotyped at the Center for Inherited Disease Research (CIDR) at Johns Hopkins University using the Illumina 660W Quad platform. Individuals and SNPs with a missing genotype rate of 2% or more were excluded. Any SNP with a minor allele frequency (MAF) less than 0.03 or failing a test of Hardy-Weinberg equilibrium with a p-value less than 1.0×10^{-5} were

also excluded. Tests of familial relationships (second degree or higher) and validation of reported sex were carried out using the GWAS data. Individuals with familial relationships or with misreported sex, that could not be resolved by reevaluation of the original records were excluded from the data. Multidimensional scaling and cluster analysis was performed to assess population stratification and admixture. A single predominant cluster was revealed and several ancestral outliers were excluded. After all QC steps, 2185 individuals, including 1668 confirmed celiac disease cases and 517 unaffected controls, were used for association analysis in the current study. 1898 SNPs between positions 26,000,508 and 33,544,122 on chromosome 6p, encompassing the xMHC region were used for the current association analysis.

4.3. Results

The xMHC SNPs were analyzed for association with celiac disease using a simple logistic regression model that included only the SNP genotype. This analysis was performed to assess the associations between the xMHC SNPs and celiac disease without any adjustment for the known high-risk genotypes at the *HLA-DQA1* and *HLA-DQB1* genes. The result of this association analysis is shown in Figure 1. The SNP with the strongest association was rs2647044 with several other SNPs also showing significant association near the *HLA-DQA1* and *HLA-DQB1* genes. SNP rs2647044 is approximately 35 kb from *HLA-DQB1* and 60 kb from *HLA-DQA1* and in nearly perfect LD with both genes. Pairwise LD analysis, as measured by the r^2 value, between rs2647044 and all other SNPs showed that there was no other SNP amongst those tested that were strongly correlated with rs2647044 (Figure 4.1). This was expected as tag SNPs chosen to be on the Illumina GWAS platform are selected to be highly informative, have low redundancy, and show low r^2 values. In figure 1, recombination hotspots are also clearly apparent and show changes in the patterns of association between the SNPs and disease (Figure 4.1).

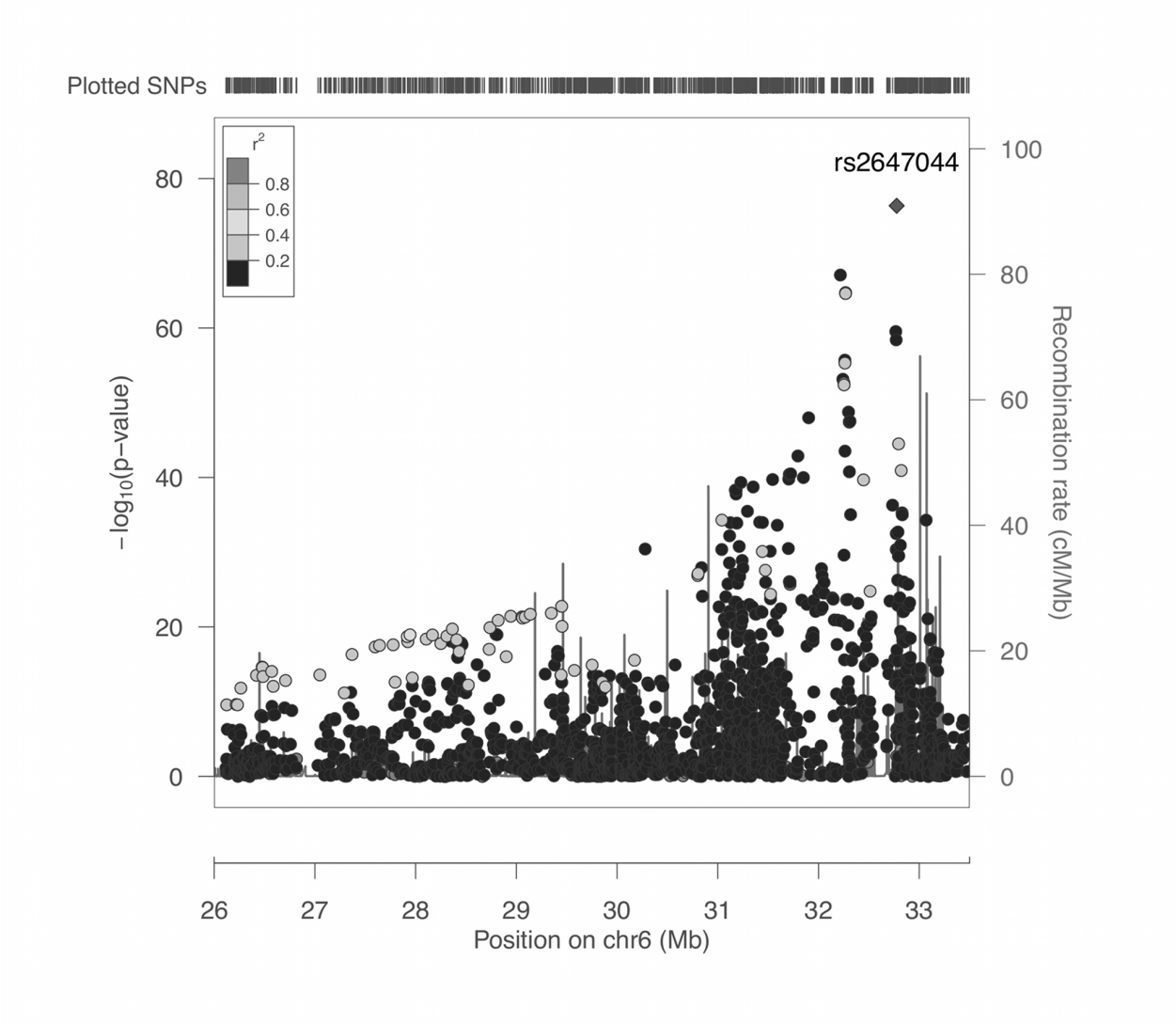
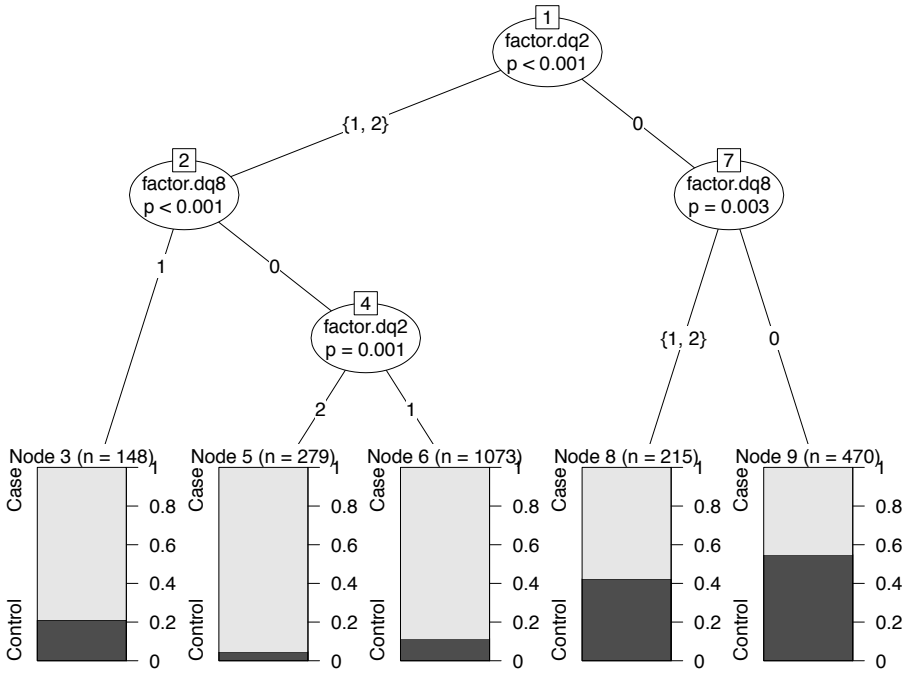


Figure 4.1. Association results for 1898 SNPs across xMHC, without accounting for the known HLA high-risk genotypes in the statistical analysis. Vertical bars indicate recombination rates generated from HapMap database. All pairwise linkage disequilibrium coefficients (r^2) included the most significantly associated SNP, rs2647044.

As the HLA high-risk genotypes made it difficult, if not impossible, to identify independent associations, a statistical procedure was implemented to generate a categorical variable to represent the known HLA haplotype effects and this categorical variable was included in the logistic regression model. The HLA genotypes that impart the highest risk of developing celiac disease are HLA-DQ2.5 and HLA-DQ8. Through in-trans combination of haplotypes, HLA-DQ2.2/7 also result in

the HLA-DQ2.5 genotype. *HLA-DQA1* and *HLA-DQB1* genotyping resulted in 16 possible multi-haplotype categories. A conditional inference tree model based on the *HLA-DQA1* and *HLA-DQB1* genotypes was computed that resulted in five terminal nodes (Figure 4.2). As was expected, the terminal node that predicted the highest proportion of celiac disease cases was the node for the DQ2.5 homozygotes (n = 279 samples), followed by heterozygote terminal node with one copy of DQ2.5 and one non-DQ8 haplotype (n = 1073), and the DQ2.5/DQ8 heterozygote node (n = 148 samples). The proportion of predicted cases dropped off significantly for the final two terminal nodes, with the worst prediction rate for those samples with both non-DQ2.5 and non-DQ8 haplotypes (n = 470). A binary split could not be made at the DQ8 terminal node because there was no discernible difference in prediction rate between samples having 1 or 2 copies of the DQ8 haplotype (n = 215). The conditional inference tree model did not provide evidence that the effects of either the DQ2.5 or the DQ8 haplotypes were multiplicative. The frequency distribution of the categorical variable created from the five terminal nodes of the recursive partitioning analysis is show in Table 1. Each binary split in the conditional inference tree was statistically significant at $P < 0.003$.



Node 3: DQ2.5/8 heterozygote
Node 5: DQ2.5 homozygote
Node 6: DQ2.5 and non-DQ8
Node 8: 1 or 2 copies of DQ8
Node 9: Non-DQ2.5 and non-DQ8

Figure 4.2. Binary tree computed by conditional recursive partitioning on HLA-DQA1 and HLA-DQB1 genotypes.

Table 4.1. Characteristics of Five-Level Variable for Known HLA High-Risk Alleles Computed by Conditional Recursive Partitioning.

Genotype	Cases	Controls	Total
DQ2.5/DQ8 heterozygote	117 (0.07)	31 (0.06)	148 (0.07)
DQ2.5 homozygote	266 (0.16)	13 (0.03)	279 (0.13)
DQ2.5/non-DQ8 heterozygote	949 (0.57)	124 (0.24)	1073 (0.49)
1 or 2 copies of DQ8	124 (0.07)	91 (0.18)	215 (0.10)
non-DQ2.5/non-DQ8	212 (0.13)	258 (0.50)	470 (0.22)
Total	1668 (1.00)	517 (1.00)	2185 (1.00)

After performing a simple logistic regression, the xMHC SNPs were tested in a multiple logistic regression model that included the HLA haplotype categorical variable as well as SNP rs1063355. SNP rs1063355 was included in the adjusted association model because it is in the 3' untranslated region (UTR) of the *HLA-DQB1* gene and is strongly associated with the *HLA-DQB1* high-risk allele. While rs1063355 was amongst the Illumina GWAS SNPs it was not one of the tagging SNPs used to determine HLA high-risk alleles. SNP rs1063355 was included in the association model to reduce the possibility of the SNP being identified as an independent predictor of the disease and to account for possible residual effects from *HLA-DQB1* that were not identified by the categorical variable. In a simple logistic regression model with only rs1063355, the SNP had a p-value of less than 1.0×10^{-17}. When the computed categorical variable was added to the model $P_{rs1063355}$ increased to 2.5×10^{-5}, indicating that rs1063355 has a relatively less significant effect but still may have significant residual effects on the disease outcome. Figure 4.3a demonstrates the impact of the adjustment for the known HLA high-risk genotypes. While rs2647044 was still very significant ($P = 3.85 \times 10^{-20}$), rs9357152 ($P = 7.28 \times 10^{-24}$) became the most significantly associated SNP in the xMHC region when the HLA high-risk genotype effects were accounted for in the multiple logistic regression model. The SNP rs9357152 is not in LD with the *HLA-DQA1* or *HLA-DQB1* genes or any other SNPs in the xMHC region. In Figure 3b, magnified on the *HLA-DQA1* and *HLA-DQB2*

intergenic region, several other SNPs (other than rs9357152) are observed to be significantly associated within the region.

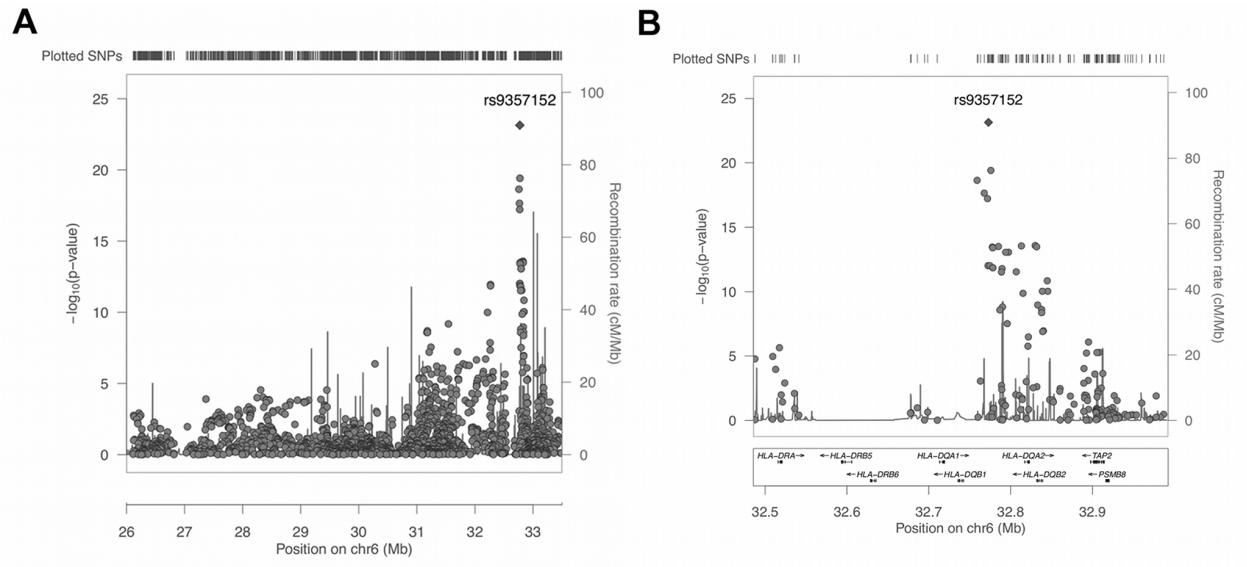


Figure 4.3. Association results for 1898 SNPs across (a) full xMHC, and (b) focused on the region around the known HLA class 2 celiac disease genes, accounting for known HLA high-risk genotypes in the statistical analysis. Vertical bars indicate recombination rates generated from HapMap database. All pairwise linkage disequilibrium coefficients (r^2) included the most significantly associated SNP, rs9357152.

A SNP fine-mapping method was carried out to identify a minimal set of SNPs that are likely to be independently associated with celiac disease across the xMHC. This fine-mapping analysis used the results from the adjusted association analysis that accounted for the known *HLA-DQA1* and *HLA-DQB1* high-risk genotype effects. These minimal, independent sets of SNPs were created according to their correlated effects on the outcome variable, with each set identified by a single most informative index SNP that was most strongly associated with disease. The following parameters were implemented for the grouping: 1) index SNP significance threshold of $P < 5x10^{-7}$; 2) grouped SNPs significance threshold of $P < 0.01$; 3) a physical distance threshold downstream

and upstream of 250 kb over which the SNPs in the group can span; 4) an LD threshold of $r^2 \geq 0.1$ with the index SNP.

While these grouping parameters were chosen to be conservative, a sensitivity analysis was performed to determine the relationships between the parameter values used in the SNP grouping procedure and the number of index SNPs identified. As described above, the SNP grouping procedure depends on several user-defined input parameters. This analysis determined the sensitivity of the grouping procedure to the input parameter values and demonstrated that parameter values used in subsequent analyses were not arbitrary. This sensitivity analysis showed that the number of index SNPs depended largely on the significance threshold value of the index SNP and the LD value between the index SNP and the secondary SNPs that are within each group of SNPs. Figure 4a shows a linear relationship between the numbers of index SNPs identified and the r^2 parameter, with r^2 values ranging from 0.05 to 1.0. As the r^2 value rises, the number of groups tended to rise while the number of secondary SNPs tended to decrease, resulting in more index SNPs. The relationship between the number of index SNPs and the r^2 values between 0.01 and 0.15 is shown in Figure 4.4b, where the number of index SNPs reaches a minimum of six at an r^2 value of 0.07. The threshold value for the r^2 value was set at 0.10 as an optimal compromise between the ability to distinguish independent associations and accepting that long-range LD is predominant across the MHC and weak correlations between independently acting loci are likely. Figure 4.4c shows the relationship between the significance threshold for the index SNP [as measured by $-\log_{10}$ (p-value for association)] and the number of index SNPs is shown in Figure 4.4c. The number of index SNPs dropped off sharply as the index SNP statistical significance threshold increased [i.e., the p-value decreased and $-\log_{10}$ (p-value) increased], with the parameter showing a less discernible impact on the number of index SNPs starting at a p-value of 1.0×10^{-6} [$-\log_{10}$ (p-value) = 6]. The number of index SNPs tends towards zero as the index SNP threshold exceeds the minimum observed p-value of all

SNPs in the experiment. The index SNP p-value threshold of 5.0×10^{-7} [$-1 \times \log_{10}$ (p-value) = 7] was selected because any value below this point has a strong effect on the number of index SNPs identified while still allowing for selection of many putatively independent loci. This sensitivity analysis determined that the secondary SNP significance level and the physical distance parameters had weak, indiscernible effects on the number of index SNPs identified (results not shown).

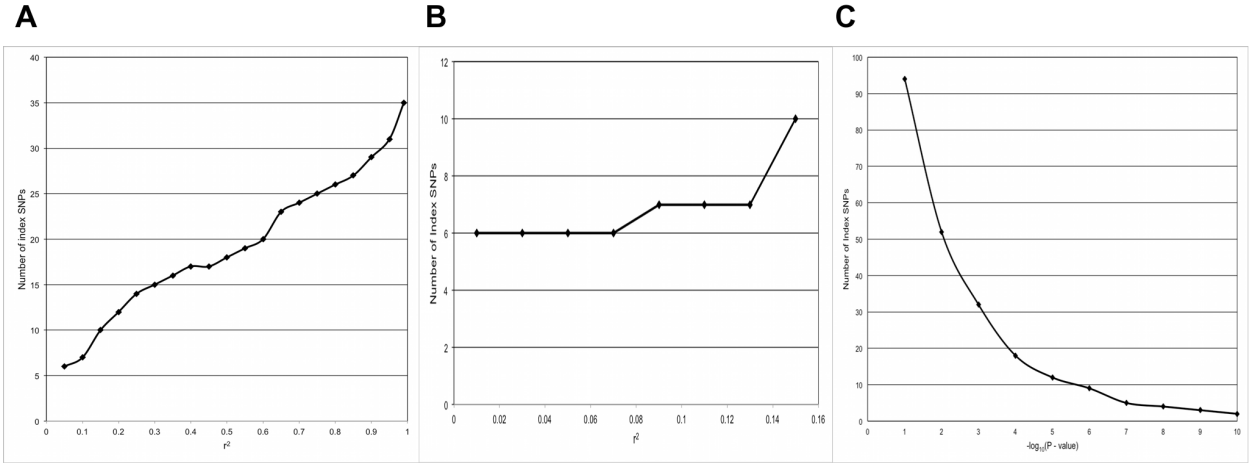


Figure 4.4. Results of sensitivity analysis for SNP grouping analysis showing the relationship between the group linkage disequilibrium parameter (r2) and the number of index SNPs identified, with r2 ranging from (a) 0.05 to 0.95, and focused on the range from (b) 0.01 to 0.15. Results (c) show the relationship between the minimum statistical significance parameter for the association between the disease and the index SNP and the number of index SNPs identified.

The fine-mapping procedure resulted in the identification of seven index SNPs in the classical MHC region associated with celiac disease, in addition to the known HLA high-risk genotypes that were accounted for in the adjusted analysis. The positions of these seven loci across the xMHC are shown along with the estimated rates of recombination in Figure 4.5, where all seven loci fall within the classical MHC region and are separated by hotspots of recombination. Table 4.2 ranks the seven index SNPs by p-value, with all seven SNPs showing $p < 5.0 \times 10^{-7}$. Four of these SNPs had odds ratios greater than 1.0 indicating that the minor allele occurred with greater

frequency among cases, while the other three SNPs showed the opposite effect direction and the major allele occurred more frequently among cases. The top four index SNPs (rs937152, rs204991, rs2523674 and rs2517485) each tagged over 30 secondary SNPs for a combined total of 135 SNPs, while the other three index SNPs (rs2260000, rs9276435 and rs2844776) combined tagged a total of 34 secondary SNPs (Table 4.2).

Further verification of the independence of the seven index SNPs was conducted by simultaneously testing SNPs for association in two multivariate logistic regression models. In the first multivariate model, only the seven index SNPs were included as predictors of celiac disease. As reported in Table 4.2, five of the seven index SNPs remain statistically significant with $p < 0.01$ when all of the index SNPs are simultaneously tested in a multivariate model. Two SNPs, rs2260000 (p-value = 0.23) and rs2844776 (p-value = 0.052), did not remain statistically significant when tested simultaneously with the other five SNPs. The second multivariate logistic regression model included the seven index SNPs as well as the five-level factor variable capturing the known common HLA high-risk genotypes. This multivariate analysis revealed that four of the seven SNPs were still statistically significant predictors of celiac disease: rs9357152 (p-value = 0.00012), rs204991 (p-value = 0.0024), rs2523674 (p-value = 0.00693) and rs2517485 (p-value = 0.0022).

Table 4.2. Association Results for Seven Index SNPs Representing Independent Loci.

SNP	Position	Minor allele	Minor allele freq.	P-value	Odds ratio	No. Secondary SNPs	Multiple LR P-value NO HRA Adj.#	Multiple LR P-value HRA Adj.	Functional gene*
rs9357152	32664960	G	0.12	7.28×10^{-24}	0.24 (0.21-0.28)	34	4.0×10^{-4}	1.2×10^{-4}	*HLA-DQB1*
rs204991	32161366	G	0.46	1.13×10^{-12}	2.24 (2.00-2.51)	33	3.31×10^{-9}	2.4×10^{-3}	*GPSM3*
rs2523674	31436789	A	0.35	6.55×10^{-10}	0.57 (0.52-0.62)	33	1.4×10^{-3}	6.93×10^{-3}	*HCP5*
rs2517485	31074101	A	0.49	2.03×10^{-9}	1.77 (1.61-1.95)	35	6.53×10^{-5}	2.2×10^{-3}	*SEEK1/PSORS1C1*
rs2260000	31593476	G	0.22	5.90×10^{-8}	0.60 (0.54-0.66)	15	0.23	0.39	*BAT2*
rs9276435	32713867	A	0.40	3.28×10^{-7}	1.82 (1.62-2.05)	5	1.4×10^{-4}	0.10	*HLA-DQA2*
rs2844776	30171827	G	0.37	4.17×10^{-7}	1.65 (1.50-1.83)	4	0.052	0.096	*TRIM26*

*Either the gene that the SNP occurs in, or the nearest gene within the same LD haplotype block as the index SNP.
#Indicates adjustment for known common high-risk alleles by inclusion of the five-level variable computed by recursive partitioning.

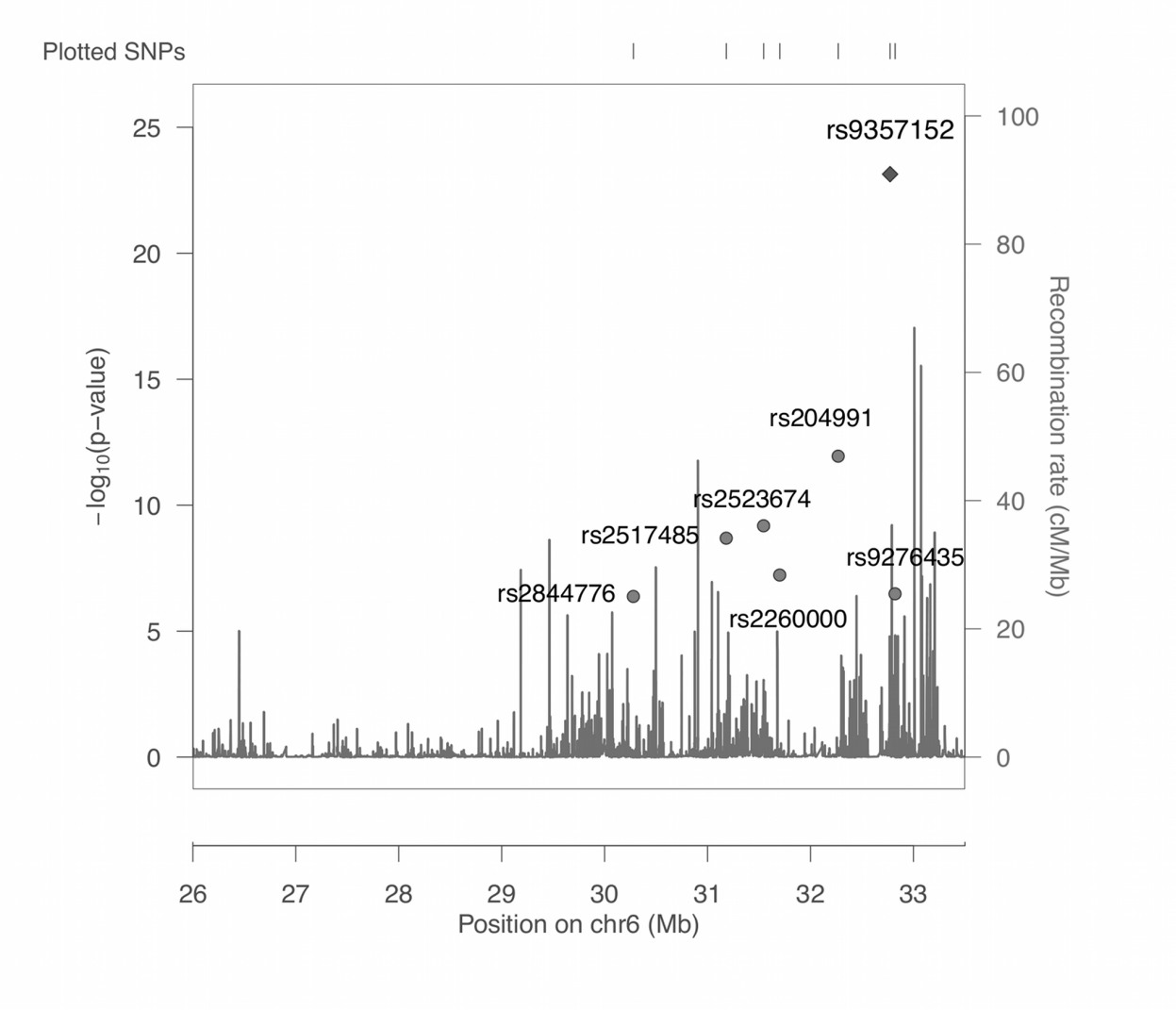


Figure 4.5. Association analysis results and locations of seven index SNPs identified by grouping analysis of the xMHC. Recombination rates were estimated from HapMap data and are indicated by vertical bars.

4.4. Discussion

GWAS of celiac disease have thus far successfully identified 39 non-HLA loci showing genome-wide significant association with celiac disease, with modest predictive information[72,73]. The high-risk alleles of the *HLA-DQA1* and *HLA-DQB1* genes may be considered necessary for development of celiac disease but are not sufficient. The xMHC region contains more than 250 expressed genes, many of which are involved in the regulation of the immune system[149]. However, it had not been

thoroughly investigated for additional celiac disease loci because of the complex nature of the analysis, which includes adjusting for the very strong effects of the known HLA disease alleles, the extraordinary genetic variation within the region, and complex patterns of linkage disequilibrium. A simple association analysis of the common variants within the region that does not take these complications into account would likely generate misleading results (as seen in the comparison of simple and multiple logistic regression models above). Statistical evidence was presented for four novel and independent celiac disease susceptibility loci within the classical MHC region. An informative measure of the known high-risk HLA genotypes was computed by conditional inference based recursive partitioning and encoded as a categorical variable that was included in an association analysis of the 7.6 Mb xMHC region using a set of 1898 SNPs that passed rigorous GWAS quality control assessments. The conditional inference based recursive partitioning approach is superior to a sample stratification to account for the known HLA high-risk types because power is not lost from sub-sampling while creating an informative measure of the effect of the haplotypes. The classification and regression trees algorithm (CART) implemented by Nejentsev et al.[140] has the potential problems of model overfitting and a bias towards selecting a model with too many binary splits because CART does not take statistical significance into account when making a binary split. The conservative fine-mapping method implemented in this study to identify the independent associations minimizes the probability of seeing false positive results.

The conditional inference based recursive partitioning analysis generated a variable that captured the effects of the known common *HLA-DQA1* and *HLA-DQB1* high-risk genotypes in a highly informative factor variable. Including this factor variable in the association analysis had a distinctly noticeable effect on the association results. Of the 671 SNPs that had $p < 5.0 \times 10^{-7}$ from the simple logistic regression analysis, only 48 SNPs had p-values less than this threshold when the known HLA effects were accounted for in the adjusted model. While this statistical adjustment likely

captured much of the known HLA effects, it is also very likely that the adjustment was incomplete. There are rare and low frequency risk alleles and genotypes in the *HLA-DQA1* and *HLA-DQB1* genes that were not specifically identified by the factor variable because they are not common enough to be strong predictors of disease in the full sample. Furthermore, the complex genetic structure of the MHC is not completely resolvable by a straightforward statistical adjustment as employed in this study. While residual influence from the *HLA-DQA1* and *HLA-DQB1* high-risk genotypes cannot be entirely ruled out, it is not likely that the four independent disease loci identified in this study are due to correlation with the known HLA high-risk genotypes given the conservative approach that was taken in the study. The results presented here strongly suggest additional loci associated with celiac disease are present in the MHC region and may be identified in a study that incorporates rare and low-frequency variants (MAF $\leq$ 0.01 and MAF $<$ 0.05, respectively) from either a custom, dense microarray chip such as the Illumina Immunochip platform or through resequencing.

Of the SNPs reported in Table 4.2, none occur in the exonic regions of genes or have a reported functional effect. Table 4.2 also lists the genes that each index SNP occurs in or if it's not known which gene the SNP belongs in, the closest gene that is within the same LD haplotype block. SNP rs9357152 was previously reported to be associated with celiac disease[68], occurs on the same haplotype block as rs9469220, a SNP reported to be associated with related autoimmune disease, Crohn's disease[150], as well as rs6457617, a SNP reported to be associated with another related autoimmune disease, rheumatoid arthritis[151]. While *HLA-DQB1* is the closest gene to rs9357152 at approximately 45 kb centromeric, a moderate amount of recombination separates this index SNP from the *HLA-DQB1* gene. SNP rs204991 is located on the third intron of the G-Protein Signaling Modulator 3 gene (GPSM3) and on the haplotype block that encompasses the entire gene. The nearest gene to rs2523674 is the gene, HLA Complex P5 (HCP5), which is about 4 kb away; there

are no other genes within 20 kb of this SNP. SNP rs2517485 is located about 10 kb from genes, SEEK1/PSORS1C1, that is implicated in susceptibility for psoriasis and systemic sclerosis disease.

The aim of this study was to test the hypothesis that the known *HLA-DQA1* and *HLA-DQB1* celiac disease high-risk alleles were not the only celiac disease alleles within the xMHC region. While the results show evidence for additional, independent celiac disease loci within the 3.7 Mb classic MHC region, no statistically significant evidence was found for additional disease loci within the additional 4.1 Mb that make up the xMHC region in the adjusted analysis. The index SNPs that were identified, SNPs rs9357152, rs204991, rs2523674 and rs2517485, were the common (MAF $\geq$ 0.05) markers with strongest evidence for association for the four new loci in the xMHC. Additional investigation in a follow-up replication study will be required to validate the reported findings and to locate additional novel disease alleles, including determining if these SNPs are causal, perhaps playing key roles in regulation of genes in the MHC region, or if they are only tagging causal variants that are not yet identified.

Chapter 5: Identification of Rare and Low-frequency Variants Associated with Celiac Disease in the 12 Previously Identified Regions and the MHC Region

Much of the heritability of celiac disease that is not attributed to the high-risk *HLA* loci has yet to be explained. Multiple GWASs have uncovered 39 non-*HLA* loci spread across 12 genomic regions on chromosomes 1, 2, 3, 4, 6, 11, 12, and 16 but explain less than 15% of the heritability of celiac disease. The purpose of this study is to use targeted resequencing data of regions previously identified to harbor common variants associated with celiac disease and identify rare and low-frequency variants that may be associated with celiac disease.

5.1. Introduction

The role of common variants in the genetic susceptibility of celiac disease (CD) has been extensively interrogated by multiple genome-wide association studies (GWAS). The first GWAS of CD in 2007 by van Heel et al.[68] identified genome-wide significant (GWS) common risk variants in the LD block that includes the genes *IL2* and *IL21* on chromosome 4q27. Hunt et al.[69] quickly followed up on this first GWAS of CD by identifying seven more risk regions associated with CD genome-wide, including regions that are known to harbor immune response genes such as *CCR3*, *IL12A*, *IL18RAP*, *RGS1*, *SH2B3*, and *TAGAP*. Another two novel regions associated with CD, 6q23.3 (*OLIG3-TNFAIP3*) and 2p16.1 (*REL*) were discovered by Trynka et al.[70] in 2009. A replication GWAS of celiac cases from the USA was performed by Garner et al.[71], confirmed the associations found in five of the eight regions that had been previously identified and provided evidence for a new candidate gene on chromosome 2q31, *ITGA4*. This was followed by a study by Dubois et al.[72] that further identified 13 more regions with GWS evidence and another 13 regions with suggestive evidence of association with CD ($p < 1x10^{-6}$). A fine-mapping study of 183 previously identified loci using a custom dense genotyping array that captures low-frequency and rare variants was carried out in 2011 by Trynka et al.[73], and while this study did not find statistically significant evidence for low-

frequency or rare variants, it did reveal evidence for another 13 novel celiac disease loci and brought the total number of non-*HLA* risk loci to 39. However, even with the identification of dozens of celiac disease risk loci, only about 14% of the heritability (excluding the *HLA* loci, which explain about 40% of the heritability) is accounted for by the non-*HLA* risk loci.

Under the common disease, common variant hypothesis (CDCV)[152], a model that posits that complex diseases can be largely attributed to common variants (minor allele frequency (MAF) ≥ 5%), each common variant discovered by a GWAS is thought to explain several percent of the population risk for a given disease. GWASs make use of an array of polymorphic markers identified through large-scale projects such as the International HapMap Project[153] and the 1000 Genomes Project[154], that are hypothesized to indirectly represent causal variants via common variants(also known as 'tag SNPs') that are in association with a disease or trait. While GWASs have yielded scores of significantly associated common variants for many common diseases, the common variants have failed to explain a meaningful percentage of the heritability. This 'missing heritability' problem[155], as it has come to be known, has driven the adoption of the common disease, rare variant model (CDRV)[156]. According to CDRV, much of the heritability of common, complex diseases is due to moderate-to-high penetrance rare variants. While CDRV recognizes the role of regulatory loci and the environment in the differential expression of a disease or trait, the model places much of the disease heritability on rare variants[17]. According to evolutionary theory, variants that are disease causing should not be common because disease tends to be deleterious to reproductive fitness and as such, disease causing variants should be selected against and prevents such variants from attaining a higher frequency in a population[157,158]. However, because selection does not remove every deleterious variant, the variants that have a more modest effect on reproductive fitness may attain a detectable population frequency and it has been argued that the assumption of purifying selection may be relaxed with humans[152,159,160]. Furthermore, data from population genetics studies

and data from whole-exome sequencing studies have demonstrated that there is an excess of rare variants, particularly nonsynonymous, deleterious variants[161–163].

Next-generation sequencing

Next-generation sequencing (NGS), in the form of high-throughput whole-exome sequencing, whole-genome sequencing, or targeted resequencing of previously identified regions, has rapidly gained traction since its inception because it has allowed researchers to directly identify potentially causal variants by genotyping all bases at a given locus, including any rare variants, at a cost, both in terms of time and money, that is acceptable. Prior to NGS, automated Sanger sequencing, now known as "first-generation" sequencing technology, was the preferred technology to perform resequencing. Indeed, initial sequencing of the human genome was performed using automated Sanger sequencing at a cost of approximately $2.7 billion and over a ten year span of time[164]. Sanger sequencing itself was first introduced by Fred Sanger in 1977[165] and except for automation and the use of fluorescent probes in place of radioactive probes, it has changed little since its introduction and is still widely considered the gold-standard for clinical cytogenetic applications. The primary strength of Sanger sequencing is its sequencing chemistry, which has been well refined over the last three decades and is still the most accurate sequencing method available and produces long reads of DNA fragments that range between 500 bases and 1 kilobase (kb) in length[166].

Sequencing throughput is a function of the number of sequencing reactions that can be run simultaneously and the lengths of the reads for each of these sequencing reactions and sequencing throughput has become the main limiting factor of Sanger sequencing. The throughput of Sanger sequencing is very limited because it requires electrophoretic separation of DNA fragments[167]. The most efficient automated Sanger sequencing machine can run only 96 sequencing reactions in parallel with a maximum output of 115 kb reads per day of operation. To deal with the problems

posed by the low throughput and high cost of Sanger sequencing, the NHGRI began funding NGS technology in 2004 with the dual goals of reducing cost to around $1,000 for a full human genome and significantly increasing throughput within a ten year timeframe[168,169].

NGS technology (also known as "massively-parallel" sequencing or "second-generation"), is actually a collection of competing sequencing technologies developed over the last decade that have dramatically increased sequencing throughput by several orders of magnitude while simultaneously decreasing the cost by orders of magnitude as well. In principle, all of the competing NGS technologies read DNA fragments that have been immobilized on some template array and differ in template generation chemistry and detection methods[167].

Before any sequences can be read using NGS technology, a sequencing library must be created. A sequencing library consists of fragmented, adapter ligated pieces of single-stranded DNA with fragment length depending on the particular technology[170]. The DNA fragments are attached to either a solid surface or to a bead. After library creation, the DNA fragments can either be clonally amplified and sequenced or sequenced directly. Clonal amplification of a sequence template is typically required in most NGS technologies for proper detection of the addition of a nucleotide[171,172]. However, DNA polymerases for template amplification do introduce mutations that can result in false positive genotype and variant calling further down the work flow. Despite the possibility of sequencing errors from amplification, the error rate is still lower overall than for NGS technologies that allow for single-molecule sequencing without clonal amplification; these single-molecule sequencing technologies are currently immature relative to technologies requiring clonal amplification and will not be discussed here.

Despite the markedly improved throughput of NGS technologies, the most widely used platforms by Illumina and Roche 454 tend to have shorter average read lengths (36 to 150-bp and ~400 to ~700-bp, respectively) than their first-generation sequencing counterparts[164,167], though the

latest line of Roche 454 based platforms claim average read lengths up to ~1-kb. These shorter read lengths do make experiments such as the assembly of a genome *de novo* very difficult[173] though not impossible. After fragmentation, the reads must be re-assembled and are typically done so by alignment back to the appropriate reference genome of the organism that is being sequenced. However, short-reads from most of the NGS instruments that are widely used do not align well to a reference genome for repetitive regions and may leave gaps in genome coverage in these regions. To aid in achieving optimal coverage of genomic regions using short-reads, a modification to the DNA library preparation allows for the generation and reading of both the forward and reverse template strands of a given short-read in what is known as paired-end sequencing.

NGS Bioinformatics

The high throughput nature of NGS technologies is one of the key benefits but has also become one of its limitations, a limitation that will probably be overcome in the years to come as informatics methods catch up to the data generation capabilities of the current NGS instruments. NGS technologies now routinely produce sequence data sets in the range of gigabases and that places a huge strain on every aspect of informatics, from data storage, quality control, annotation, variant calling, and interpretation of the data[174]. With Sanger sequencing, it was thought that data generation was the bottleneck; with NGS technologies the opposite has now become true with NGS instruments outpacing the development of innovative informatics methods to mine and interpret these huge data sets[175]. One of the first steps after data generation and base calling in an NGS work flow is the alignment of the short reads to a reference genome (unless *de novo* assembly is to be attempted), an area of bioinformatics research that is still actively producing new methods to allow for more efficient and accurate alignment of reads to a reference genome[176]. Many of the current crop of short read alignment algorithms are based around the "Burrows-Wheeler Transform" (BWT) compression algorithm[177] or some form of hashing. Three of the most widely used alignment

packages are all based on the BWT algorithm: Bowtie[178], SOAP2[179], and BWA[86,180]. All three of these methods are very fast and memory efficient and have high sensitivity. However, hash-based algorithms such as Novoalign[181] are still slightly more sensitive, particularly in aligning repetitive reads although these marginal sensitivity gains come at a high cost in terms of memory and time. For genomic regions with high levels of variation, alignment tends to be more error prone. In particular, the alignment of highly diverse regions such as the MHC region on chromosome 6 has remained a challenge. Some of the solutions suggested are the usage of paired-end or mated-end reads and the usage of longer reads. There is also active research in hybrid *de novo* assembly/alignment to reference approaches for assembling the reads for the MHC; these approaches will not be explored in the present study.

After alignment of short reads, it is often necessary to recalibrate the Phred-scaled quality scores for each short-read[182]. The Phred score is given by the equation, $Q_{Phred} = -10 \log_{10} P(\text{error})$, such that a 1% error rate in base calling equals a Phred score of 20. This recalibration must be done for accurate downstream analysis because the raw Phred score that is generated by the proprietary base-calling algorithms that are used in conjunction with a given NGS instrument may not be accurately reporting the true error rate[183]. As mentioned above, the Phred scores must be well calibrated because the genotype and variant calling is highly dependent on the base quality score. One widely used algorithm for recalibrating Phred scores has been implemented in the widely used variant calling package, GATK[85], and takes the machine cycle and dinucleotide context into account when recalibrating the Phred scores. This recalibration algorithm operates by first estimating the empirical Phred score with respect to the reference genome and then estimating the recalibrated score by finding the difference between the empirical quality scores and the raw quality scores.

Statistical tests for low-frequency and rare variants

In the last few years, NGS technology has delivered on the promise to provide a wealth of genotypic data that has long been sought after by researchers. However, there is substantial recognition that the ability to quickly and cheaply generate genotype data has outpaced the research community's ability to interpret the data well. For instance, while statistical significance thresholds have more-or-less been agreed upon for GWASs (the now conventional $p < 5x10^{-8}$ threshold for arrays having $1x10^{6}$ SNPs), there is a paucity of simple statistical guidelines that are widely accepted for NGS-based studies[184]. The primary reason that simple statistical significance thresholds have not been as widely accepted is the different types of variants (i.e. nonsynonymous, synonymous, exonic, intronic, etc.) and their unknown prior probabilities in how they affect a disease or trait.

As locus and allelic heterogeneity are high and variants of large effect for complex diseases are likely to be low-frequency or rare, a central challenge has been the development of statistical tests that are well powered to detect low-frequency or rare variants. To identify rare and low-frequency variants associated with disease, three classes of tests have been developed: single-marker tests, multi-marker tests, and collapsing methods. Of these, the simplest and the least powerful is the single-marker class of tests. With case-control data, single-marker tests include the chi-square test, Fisher's exact test, the Cochran-Armitage (CA) test for trend, and logistic regression. The CA test for trend assumes that the genotypes are ordered and tests for a linear trend in proportions that are weighted by the number of alleles[185]. The chi-square and Fisher exact tests test the null hypothesis of no difference in the genotype frequencies at a given variant between cases and controls with the key difference being that the Fisher's exact test yields exact results and should be used when any genotype count is less than 5 in either comparison group, which is often the case with rare variants. In any case, even without the reduction of power that comes from adjustment for multiple comparisons (e.g., Bonferroni correction) across each variant, the markedly reduced power of single-marker tests to detect rare and low-frequency variants associated with disease have been shown to

be very sensitive to sample size; extremely large sample sizes of tens of thousands of samples are required to detect association signals from rare or low-frequency variants[185].

Multiple marker methods that tests all variants simultaneously can be implemented using a multivariate test such as Fisher's method or Hotelling's T^2 test. A multiple marker test has higher power to detect an association between a rare variant and a phenotype than a single-marker test unless there is only one clearly associated variant in the tested region or unit of interest. Perhaps the most simple multiple marker method, Fisher's method combines single-marker tests across all variants and has the following test statistic,

$$X^2 = -2\sum_{i=1}^{m} \log(p_i),$$

where p_i is the ith p-value from m single-marker tests. Given that H_0 is true and all p_i are truly independent of each other, then X^2 follows a χ^2 distribution with $2m$ degrees of freedom. To implement Hotelling's T^2 test, indicator variables for case and control genotypes for variants across each individual must first be defined. The indicator variable for cases, X_{ij}, is defined,

$$X_{ij} = \begin{cases} 1 \text{ if genotype is } AA \\ 0 \text{ if genotype is } Aa \\ -1 \text{ if genotype is } aa \text{ (wild-type)} \end{cases},$$

for the ith individual and jth variant. The indicator variable for controls, Y_{ij}, is defined similarly.

Then, $\overline{X}_j = \frac{1}{N_A \sum_{i=1}^{N_A} X_{ij}}$, $\overline{Y}_j = \frac{1}{N_{\bar{A}} \sum_{i=1}^{N_{\bar{A}}} Y_{ij}}$ and $\overline{X} = (\overline{X}_1,...,\overline{X}_m)^T$, $\overline{Y} = (\overline{Y}_1,...,\overline{Y}_M)^T$ where N_A is the number of cases, $N_{\bar{A}}$ is the number of controls, $\overline{X}$ is the vector of expected indicator values across cases, $\overline{Y}$ is the vector of expected indicator values for controls. Hotelling's T^2 test has the following test statistic,

$$T^2 = \frac{N_A N_{\bar{A}}}{N_A + N_{\bar{A}}} (\overline{X} - \overline{Y})^T S (\overline{X} - \overline{Y}),$$

where S is the covariance matrix of X and Y. Under H_0, T^2 follows $F_{m,N_A+N_{\bar{A}}-m-1}$. Although multi-marker tests are in general more powerful than single-marker tests, they are still sensitive to minor allele frequencies like single-marker tests. Furthermore, a simulation study by Li and colleagues[186] has shown that the multiple marker method using Hotelling's T^2 test has significantly reduced statistical power as the number of rare causal variants increases.

Collapsing methods (also known as "burden tests", in reference to the genetic "burden" that is due to rare variants) may have the most power to detect an association between disease and rare variants because these methods aggregate or collapse low-frequency and rare variants before testing for association, and in aggregate form, low-frequency and rare variants may actually be common. Collapsing approaches apply either a univariate test (i.e. the Pearson χ^2 test) or a multivariate test to the aggregated or collapsed variants within a defined group (i.e. a gene or pathway), and have the increased power of a multi-marker test while avoiding the penalty of high degrees of freedom from a multivariate analysis of individual variants or the reduction in power from multiple comparisons that a single-marker test faces. For a univariate collapsing method, indicator variables for cases and controls, X_i and Y_i respectively, must first be defined that indicate whether a rare variant is either present or not across all genotypes in each of i individuals,

$$X_i, Y_i = \begin{cases} 1 \text{ if a rare variant is present} \\ 0 \text{ if a rare variant is not present} \end{cases}.$$

Letting ϕ_A and $\phi_{\bar{A}}$ represent the rare variant frequency for cases and controls, respectively, the Pearson χ^2 test can be applied to test H_o: $\phi_A = \phi_{\bar{A}}$ with the following non-centrality parameter for the noncentral χ^2_1,

$$\nu_c = N\left[\frac{(\phi_A - \phi_{\bar{A}})^2}{\phi_A + \phi_{\bar{A}}} + \frac{(\phi_A - \phi_{\bar{A}})^2}{2 - \phi_A - \phi_{\bar{A}}}\right].$$

Taking advantage of the increased power of a multivariate test, the combined multivariate and collapsing method (CMC) of Li and Leal[186] was one of the first proposed collapsing methods that implemented a multivariate test such as Hotelling's T^2 test. The implementation of Hotelling's T^2 for CMC is the same as above except for the collapsing of n variants across k locus groups and each individual is represented by a k-vector of indicator variables for cases ($i = 1, \ldots, N_A$), $X_i = (X_{i1}, \ldots, X_{ik})^T$, and for controls ($i = 1, \ldots, N_{\bar{A}}$), $Y_i = (Y_{i1}, \ldots, Y_{ik})^T$. Now, under H_0, T^2 follows the $F_{k,N_A+N_{\bar{A}}-k-1}$ distribution, the k groups replacing the m markers. While CMC has a relatively high power to detect an association compared to multivariate tests, there is still a decrease of power when the number of noncausal variants is increased.

A collapsing method based on a logistic regression framework was proposed by Morris and Zeggini[187] that tests for either the presence or absence of at least one minor allele at any low-frequency or rare variant. In this model, one must first sum all of the rare variants across a particular gene,

$$X = \sum_{i}^{N} x_i ,$$

where X_i takes the value of 1 for each i^{th} rare variant. After this collapsed variant is created, a logistic regression model may be fitted,

$$\text{Logit Pr}(Y_i = 1) = \beta_0 + X\beta + \varepsilon,$$

and the following null hypothesis is tested: $\beta = 0$. The p-values are drawn from an asymptotic normal distribution of the Wald statistic. The primary limitation of the burden test approach is the assumption that all of the aggregated variants in a given gene act in one direction only (that is, the variants are either all deleterious or all protective). For genes that harbor deleterious, neutral, and protective rare variants that are associated with disease, this assumption of uni-directionality does not always hold well.

One of the many approaches that were developed to deal with the weakness of the directionality assumption is the C-alpha test developed by Neale et al.[188]. The C-alpha test works under the assumption that genes harbor a mix of deleterious, protective, and neutral rare variants in cases and controls and is able to maintain statistical power in the presence of deleterious, neutral, and protective rare variants by testing the variance of observed counts of a given rare variant against the expected variance of the given rare variant. The test statistic for C-alpha is,

$$T = \sum_{i=1}^{m} [(y_i - n_i p_0)^2 - n_i p_0 (1 - p_0)],$$

where n_i is the number of cases and controls for the i^{th} variant, y_i is the number of cases for the i^{th} variant and p_0 is a common probability of seeing n_i under the null hypothesis of no association. The variance of T is as follows,

$$c = \sum_{n=2}^{\max n} m(n) \sum_{u=0}^{n} [(u - np_0)^2 - np_0(1 - p_0)]^2 f(u \mid n, p_0) ,$$

where $m(n)$ is the number of variants with count n and $f(u \mid n, p_0)$ is the probability of observing u copies of the i^{th} variants under the null. With T and c, one can simply estimate a Z statistic that follows an asymptotic $N(0,1)$ under the null hypothesis of no association. Empirical p-values may be obtained by permuting case and control status.

There is still some frustration in the field due to the lack of agreement and clarity in statistical guidelines for NGS-based studies that is partially alleviated by the gene-based burden test because the number of tests that are performed simultaneously is based on the number of genes being tested, a number which is orders of magnitude lower than the number of total variants genome-wide. As such, even if the conservative Bonferroni correction is applied for all genes, the statistical threshold is relaxed several-fold. For smaller studies like the present study, permutation

based testing may also be implemented in to estimate empirical *p*-values because asymptotic *p*-values may be artificially high in smaller studies[189].

Recent NGS studies

In a recent large-scale NGS-based study involving the targeted resequencing of 202 genes in 14,002 people, Nelson et al.[190] found that 95% of all variants in the sample were rare (MAF ≤ 0.5%), that nearly 75% of the variants were only observed in 1-2 individuals, and that about 90% of the rare variants detected were novel (i.e. not reported in a previous study or database). The NHLBI Exome Sequencing Project[191], found that nearly 90% of variants across (and nearly 75% of nonsynonymous variants) 15,336 genes were of recent origin. These large-scale sequencing studies provided the empirical evidence of the excess of rare variants predicted by population genetics theory and also provided evidence that a large proportion of rare variants are unique to a given sample set and that detection of rare variants would be relatively difficult in smaller disease-specific case-control studies. This has been reflected in the relative dearth of strong evidence for rare variants of large effect size in autoimmune disorders. An early study by Nejentsev et al.[192] found significant evidence for association of nonsynonymous rare variants with type 1 diabetes in just one gene (*IFIH1*) in a resequencing of 144 regions identified by GWAS. In a resequencing study of inflammatory bowel disease by Momozawa et al.[193] in which 70 candidate genes were resequenced, only low-frequency nonsynonymous variants were identified in *IL23R*. To counter the loss of power from the small sample sizes of resequencing studies of autoimmune disorders, a very recent, large-scale study of six autoimmune disorders (including celiac disease) by Hunt et al.[75] that performed NGS-based exon sequencing of 25 genes in 41,911 individuals from the UK, demonstrated that rare coding variants contributed much less of the unexplained heritability than expected and claimed that 'missing heritability' of autoimmune disorders will probably not be explained by CDRV.

The present study attempted to identify candidate rare and low-frequency variants within genes associated with celiac disease from 13 regions (including the MHC region of chromosome 6) for future imputation and meta-analysis using a large GWAS collection from Dubois et al.[72]. This present study did not focus on definitive identification of individual rare or low-frequency variants associated with celiac disease because *a priori*, the present study, with ~500 samples, is not sufficiently powered to detect individual rare or low-frequency variants. The present study is distinguished from the Hunt et al.[75] study because non-coding regions (e.g. intronic variants) have also been resequenced and will also be interrogated.

5.2. Methods

Samples

From the 2,300 samples collected for a previous study by the North American Celiac Disease Consortium[74], 250 celiac disease cases, 239 healthy controls of Caucasian ancestry were selected to be resequenced. Signed informed consent forms for all samples along with approval from each respective Institutional Review Board was obtained and detailed in the previous study.

Sequencing

100 bp paired-end reads from targeted sequences were obtained for 509 individuals. For 106 samples, the sequencing libraries were enriched with Agilent SureSelect baits and 403 samples were enriched with Roche Nimblegen baits that cover 64 genes in 12 regions (table 5.1) across chromosomes 1, 2, 3, 4, 6 (non-MHC), 11, 12, and 16 previously found to be associated with celiac disease[68,69,72,73]. The MHC region (chromosome 6: 28.7 Mb – 33.5 Mb), with 202 genes, was covered only in the 403 samples that were enriched with the Roche Nimblegen baiting platform. All samples were sequenced at an average read depth of 35x on an Illumina GAIIx sequencing platform. To optimize variant calling, it has been shown that an average depth of 40x is desirable[194]; the present

study has an average depth, genome-wide of 35x, comparable to the average read depth reported in a recent large-scale study by Hunt et al.[75]. As such, depth of coverage for the present study is close to optimal for accurately detecting variant calls.

Table 5.1. Twelve non-MHC regions that were resequenced and genes in those regions.

Chromosome	Position (Mb)	Genes
1	192.4 – 192.6	*RGS1, RGS13*
2	60.9 – 61.9	*PAPOLG, FLJ16341, REL, PUS10, PEX13, KIAA1841, LOC339803, C2orf74, AHSA2, USP34, XPO1*
	181.7 – 182.5	*UBE2E3, ITGA4, CERKL*
3	69.1 – 69.5	*ARL6IP5, LMOD3, FRMD4B*
	159.5 – 159.8	*SCHIP1, IL12A*
	187.8 – 188.7	*LPP-AS2, LPP, FLJ42393*
4	122.9 – 123.6	*KIAA1109, ADAD1, IL2, IL21*
6	127.9 – 128.9	*C6orf58, THEMIS, PTPRK*
	159.3 – 159.6	*OSTCP1, C6orf99, RSPH3, TAGAP*
11	128.3 – 128.5	*ETS1*
12	111.7 – 113.1	*CUX2, FAM109A, SH2B3, ATXN2, BRAP, ACAD10, ALDH2, MAPKAPK5-AS1, MAPKAPK, ADAM1A, TMEM116, ERP29, NAA25, TRAFD1, HECTD4, RPL6, PTPN11*
16	10.9 – 11.5	*TVP23A, CIITA, DEXI, CLEC16A, SOCS1, TNP2, PRM3, PRM2, PRM1, RMI2*

Bioinformatics

The BWA algorithm[86] was used to perform mapping of the 100-bp paired-end reads from the 12 targeted regions while the recently introduced BWA-MEM algorithm[180], optimized for reads that are 100-bp and greater and more robust to sequencing error by switching between local and end-to-end alignment, was used to perform the mapping for the MHC region. The GRCh37/b37

(hg19) reference human genome[1], which includes the 6 alternate MHC haplotypes, was used as the reference genome for mapping all reads. After all reads were mapped, Picard and Samtools[87] were used to reorder and sort the SAM/BAM files, mark duplicate reads, and synchronize mate-pair information between paired-end reads. Local realignment around indels and recalibration of base quality scores was performed before calling variants using the GATK UnifiedGenotyper[85]. Variants were filtered for quality, where only variants called with a Q score of 20 or greater were emitted. All variants were annotated using the Variant Tools package[195] and the following databases: dbSNP 138[196], refGene[197], and dbNSFP[198]. Any variants that were not in dbSNP 138, refGene, or dbNSFP were considered novel and analyzed separately.

Statistical analysis

To analyze low-frequency and rare variants, two collapsing type methods were implemented, the fixed threshold burden test of Morris et al.[187] and the C-alpha test of Neale et al.[188]. For the fixed threshold burden test, *p*-values were asymptotically obtained after estimation of the regression coefficients. Under the C-alpha test, empirical *p*-values were obtained by permuting case and control status under the null hypothesis assumption that case and control status are swappable. As approximately 260 genes were tested, a *p*-value less than or equal to around $2x10^{-4}$ would indicate a significant finding (after Bonferroni correction given $\alpha = 0.05$). To obtain empirical *p*-values at least this low, 10000 permutations were performed for each gene. In both tests, variants were grouped by gene and for both tests, fine-scale QC was performed wherein only variants within genes with less than 1% missing genotypes and less than 1% missing samples were tested. For either test, to test for low-frequency variants a MAF cutoff of 0.04 was set. To test for rare variants, only variants with MAF < 0.01 were included in the association tests. These MAF cutoffs were based on MAF definitions of low-frequency ($0.01 \leq MAF < 0.05$) and rare ($MAF < 0.01$) found in the literature[75,199–201]. Variants and individuals with missing data were excluded during association testing

and not prior because missing data tends to occur non-randomly across the genome[195]. The burden and C-alpha tests were also performed with datasets that included only the known (i.e. in dbSNP 138) non-synonymous variants from either the 12 non-MHC regions or the MHC region, only the novel (i.e. non-dbSNP 138) variants from the 12 non-MHC regions, only the novel non-synonymous variants from the 12 non-MHC regions, and all non-synonymous (i.e. dbSNP 138 and novel) variants from the 12 non-MHC regions.

5.3. Results

For the 12 non-MHC regions, samples that were enriched on the Agilent platform had a median successfully mapped read rate of 97.3% while samples that were enriched using the Nimblegen platform had a median successfully mapped read rate of 97.8%. The median successfully mapped read rate for the MHC region was 99.75%. This marginally higher median map rate for the MHC region may be due to the usage of the optimized BWA-MEM algorithm[180]. The median GC content was 40% for the 12 non-MHC regions while the median GC content for the MHC region was 43% indicating little evidence of GC-content bias. After calling variants, the transition-to-transversion (Ti/Tv) ratio was estimated and for the 12 regions the Ti/Tv ratio was 1.95 and for the MHC region the Ti/Tv ratio was 1.98 which is very close to the ~2:1 ratio across the human genome[202]. To minimize testing erroneous variant calls, genotypes with an average read depth across samples less than 35 were excluded. A total of 21,061 variants were successfully called from the 12 non-MHC regions (intersection of both baiting platforms) while 49,162 variants were called from the MHC region. For the variants called from the 12 non-MHC regions, the Nimblegen baited samples had 18,018 variants while the Agilent baited samples had 17,030 variants confirming a report that the Nimblegen baited samples yielded more variants [personal communication]. For the 12 non-MHC regions, 8,742 variants were flagged as non-coding variants (41.5%), while 13,601 (27.7%) variants

from the MHC region were non-coding (table 4.2). There were 1,616 novel (non-dbSNP 138) variants from the 12 non-MHC regions and of these variants, 1,332 (82.4%) were annotated with function (table 5.3); there were no novel variants from the MHC region. Of the 21,061 variants from the 12 non-MHC regions, there were 3 short insertions and 26 deletions. There was only one short insertion among the 49,162 variants called from the MHC region and no deletions. A large proportion of the variants that were called could not be annotated because their function and/or gene is either unknown or not validated at the present time.

Table 5.2. Count of variants by type from the 12 non-MHC and the MHC regions.

	Type	**non-MHC**	**MHC**
Non-coding	Intronic	8550	12398
	3'-UTR	169	947
	5'-UTR	21	231
	Splice site	2	25
Coding	Synonymous	127	658
	Frameshift	0	8
	Missense	187	1107
	Nonsense	4	33

Table 5.3. Count of novel (non-dbSNP) variants from the 12 non-MHC regions.

Non-coding	**Intronic**	**56**
	3'-UTR	6
	5'-UTR	0
	Splice site	56
	Intergenic	1
Coding	Synonymous	71
	Frameshift	0
	Missense	1091
	Nonsense	51

Under the rare variant burden test of the genes from the 12 non-MHC regions, there were two signals that provided suggestive evidence of association, genes *RSPH3* (chr. 6:159.39-159.42 Mb, $p = 2.65 \times 10^{-3}$, 20/140 variants with MAF < 0.01) and *PAPOLG* (chr. 2:60.98-61.03 Mb, $p =$

$6.0x10^{-3}$, 17/43 variants with MAF < 0.01). When the MAF threshold was increased to 0.04 to include low-frequency variants, *TAGAP* (chr. 6:159.46-159.47 Mb, $p = 9.19x10^{-4}$, 27/64 variants with MAF ≤ 0.04) provided suggestive evidence of association. Rare and low-frequency burden tests of association that included only the novel variants, non-synonymous novel variants, and all non-synonymous variants did not yield any evidence for association. When a C-alpha test was implemented for rare variants from the 12 non-MHC regions, two genes provided suggestive evidence of association, *ARL6IP5* (chr. 3:69.13-69.16 Mb, $p = 5.0x10^{-3}$, 25/57 variants with MAF < 0.01) and *LMOD3* ($p = 5.0x10^{-3}$, 63/83 variants with MAF < 0.01). *LMOD3* (chr. 3:69.16-69.17 Mb, $p = 5.99x10^{-3}$, 70/83 variants with MAF ≤ 0.04) was still suggestive when the MAF threshold for the C-alpha test was increased to allow for the inclusion of low-frequency variants. A rare C-alpha test of all non-synonymous variants (dbSNP 138 and novel) provided evidence of an association with gene *KIAA1109* ($p = 1.0x10^{-4}$, 256/265) while a low-frequency C-alpha test of all non-synonymous variants did not yield any evidence of association with disease.

In the MHC region, the rare variant and low-frequency burden tests did not yield any genes with even suggestive evidence of association. However, under the C-alpha test of rare variants from the MHC region, gene *ABCF1* ($p = 3.0x10^{-4}$, 35/44 variants with MAF < 0.01) provided significant evidence while genes *MRPS18B* ($p = 8.0x10^{-4}$, 20/23 variants with MAF < 0.01) and *NOTCH4* ($p = 9.0x10^{-4}$, 22/81 variants with MAF < 0.01) provided suggestive evidence of association. When low-frequency variants were included in the C-alpha test of MHC genes, 10 genes were significant at $p < 1x10^{-4}$ while another 8 genes had *p*-values ranging between $2\text{-}4x10^{-4}$ (Table 5.4). As was the case with the rare and low-frequency burden tests of association for the 12 non-MHC region genes with only non-synonymous variants (dbSNP 138), both rare and low-frequency burden tests of association did not yield any statistically significant or even suggestive evidence of association. When a rare C-alpha test was performed on only the non-synonymous variants in the MHC region, gene *BTNL2* ($p <$

$1x10^{-4}$, 6/9 variants with MAF < 0.01) was found to be significantly associated with disease. When the MAF threshold for the C-alpha test of non-synonymous variants in the MHC region was increased to 0.04, 4 genes were found to be associated with disease with statistically significant evidence (table 5.5). One of these genes, *BTNL2*, was detected in the rare C-alpha test and the results did not change when the MAF threshold was increased. The other three genes that yielded significant evidence of association under the low-frequency C-alpha test were *COL11A2*, *PRRC2A*, and *TRIM40*.

Table 5.4. Results of C-alpha test of MHC genes (all variants) with MAF ≤ 0.04.

Gene	Variants tested/total variants	$P_{C\text{-}alpha}$
AGPAT1	2/12	$<1x10^{-4}$
AIF1	2/5	$<1x10^{-4}$
ATF6B	5/39	$<1x10^{-4}$
DDX39B	13/61	$<1x10^{-4}$
C6orf10	84/597	$<1x10^{-4}$
PPT2	3/14	$<1x10^{-4}$
PPT2-EGFL8	8/33	$<1x10^{-4}$
RNF5	1/10	$<1x10^{-4}$
TNXB	34/186	$<1x10^{-4}$
MICA	92/277	$<1x10^{-4}$
BTNL2	9/205	$2.0x10^{-4}$
ATP6V1G2-DDX39B	17/76	$2.0x10^{-4}$
EGFL8	4/13	$3.0x10^{-4}$
PBX2	3/11	$3.0x10^{-4}$
PRRC2A	32/63	$3.0x10^{-4}$
MUC22	32/230	$3.0x10^{-4}$
ATP6V1G2	3/9	$4.0x10^{-4}$
GPANK1	4/16	$4.0x10^{-4}$

Table 5.5. Results of C-alpha test of only non-synonymous variants in MHC genes with MAF ≤ 0.04.

Gene	Variants tested/total variants	$P_{C\text{-}alpha}$
BTNL2	6/9	$<1.0\times10^{-4}$
COL11A2	7/8	8.0×10^{-4}
PRRC2A	16/19	7.0×10^{-4}
TRIM40	3/4	8.0×10^{-4}

5.4. Discussion

This study provides statistically significant evidence that there are rare and low-frequency variants associated with celiac disease. As this study did not re-sequence only the coding regions as has been done in a previous study[75], non-coding region variants such as intronic variants, have been discovered to be associated with celiac disease. While this study in itself does not provide strong evidence to support CDRV, it has provided the identification of rare and low-frequency variants within regions that previously yielded common variants that may either be causal or involved in the expression of other genes that may be associated with celiac disease and candidates to be imputed and re-sequenced (or genotyped on a dense custom microarray) in a future study that has a much larger sample size, preferably in the tens of thousands. Regarding sample size, the present study, with approximately 500 cases and controls, is likely underpowered to detect rare and low-frequency gene-level associations across the whole genome. However, this study has shown that collapsing-based testing approaches such as the burden test and the C-alpha test are adequately powered to detect disease associations at the gene-level when considering only the candidate regions that were re-sequenced (the 12 non-MHC and the MHC).

Under a rare burden test, genes *RSPH3* (Radial Spoke 3 Homolog) and *PAPOLG* (Poly(A) Polymerase Gamma) provided suggestive evidence of association with celiac disease. It is unclear what functional role that variants in either of these genes play in etiology of celiac disease and neither of these genes have any variants that have been previously associated with celiac disease, although *PAPOLG* has been previously found to be associated with a related autoimmune disorder, Crohn's disease, in an Ashkenazi Jewish population[203]. However, *PAPOLG* is ~200kb downstream of and on the same LD-block on chromosome 2 as a gene known to harbor a common variant discovered to be associated with celiac disease, *REL* (V-Rel Avian Reticuloendotheliosis Viral Oncogene Homolog)[72]. Gene *RSPH3* is also located downstream of and on the same LD-block on chromosome 6 as *TAGAP* (T-Cell Activation RhoGTPase Activating Protein), an immune response gene that harbors a common variant previously found to be associated with celiac disease[69]. When the MAF threshold for the burden test was raised to 0.04, *TAGAP* itself also provided suggestive evidence of association with celiac disease, supporting the hypothesis that genes with common variants associated with disease may also harbor rare and low-frequency variants associated with disease[155]. Further investigation needs to be conducted to determine if the associations from *PAPOLG* and *RSPH3* are independent of *REL* and *TAGAP*, respectively, or if the rare and low-frequency variants are in LD with the known common variants.

The C-alpha test of rare and low-frequency variants in both the 12 non-MHC regions and the MHC region yielded over a dozen signals with suggestive or significant evidence of association with celiac disease. Within the 12 non-MHC regions, the rare C-alpha test yielded suggestive evidence of association for two genes, *ARL6IP5* (ADP-Ribosylation-Like Factor 6 Interacting Protein 5) and *LMOD3* (Leiomodin 3), both in the same LD-block on chromosome 3; *LMOD3* still provided suggestive evidence of association after the MAF was raised to 0.04. While neither of these genes harbor variants that have been previously found to be associated with celiac disease, both

genes are less than 50kb downstream of a gene previously found to harbor a common variant associated with celiac disease, *FRMD4B* (FERM Domain Containing 4B)[72]. Further investigation of the LD between *ARL6IP5* and *LMOD3* and *FRMD4B* is required to ascertain whether the signals that have been detected in this study are independent. The expression of *ARL6IP5* is known to be upregulated by retinoic acid and retinoic acid as a co-adjuvant with *IL-15* has been shown to be a promoter of inflammation in mouse models of celiac disease[204], suggesting a pathway to further investigate in future studies. When only all rare non-synonymous variants (dbSNP and novel) in the 12 non-MHC regions were tested under C-alpha, a previously implicated gene with a highly-significant common variant, *KIAA1109*[73], provided significant evidence of association with celiac disease, perhaps suggesting that the non-coding rare variants do not play a significant role in the development of celiac disease.

Within the MHC region, the rare C-alpha test provided suggestive evidence of association for three genes, *ABCF1* (ATP-Binding Cassette, Sub-Family F, Member 1), *MRPS18B* (Mitochondrial Ribosomal Protein S18B), and *NOTCH4* (Notch 4). *ABCF1* and *MRPS18B* are within ~20kb of each other and are both associated with other autoimmune disorders (autoimmune pancreatitis and systemic lupus erythematosus, respectively). Notably, *NOTCH4* is adjacent to the gene, *GPSM3* (G-Protein Signaling Modulator 3), that harbors a common variant, rs204991, that was discovered to be associated with celiac disease in a previous fine-mapping study[74]. The low-frequency C-alpha test of the MHC region yielded 10 genes with significant evidence and a further 8 genes with suggestive evidence of association. Of the 10 genes with significant evidence of association, *MICA* (MHC Class I Polypeptide-related Sequence A) was the only gene that has been previously associated with celiac disease[205,206]. *MICA* is a gene that expresses HLA class I molecules that are overexpressed in epithelial cells in the small intestine in response to gliadin peptides and upregulates a natural killer cell receptor (NKG2D) on intra-epithelial lymphocytes that leads to

enterocyte death. Several of the other genes bearing significant or suggestive evidence of association have been found to be associated with other autoimmune or inflammatory diseases: *DDX39B* (DEAD Box Polypeptide 39B) and *PRRC2A* (Proline-Rich Coiled-Coil 2A) with rheumatoid arthritis; *EGFL8* (Epidermal Growth Factor-Like-Domain, Multiple 8), *GPANK1* (G Patch Domain and Ankyrin Repeats 1), and *MUC22* (Mucin 22) with systemic lupus erythematosus; *C6orf10* (Chromosome 6 Open Reading Frame 10) with psoriasis; *PRRC2A* with insulin-dependent diabetes mellitus (type 1 diabetes); and *BTNL2* (Butyrophilin-Like 2) with sarcoidosis, a rare inflammatory disorder. When only the non-synonymous variants are included in the rare and low-frequency C-alpha tests, *BTNL2* and *PRRC2A* still provide significant or suggestive evidence of association. Two additional genes, *TRIM40* (Tripartite Motif Containing 40) and *COL11A2* (Collagen, Type XI, Alpha 2) provide suggestive evidence of association. *TRIM40* is known to be associated with systemic lupus erythematosus while *COL11A2* is located very close (~1kb) to *RXRB* (Retinoid X Receptor, Beta), a gene that is involved in regulating the effects of retinoic acid.

The performance of the C-alpha test relative to the burden test in both the 12 non-MHC regions and the MHC region suggest that the rare and low-frequency variants in the genes that provided significant or suggestive evidence of association are more likely to not be unidirectional with respect to effect, especially in the MHC region. While the burden test was not able to detect even weak suggestive evidence of association for MHC region genes with low-frequency or rare variants, the C-alpha test was able detect significant evidence of association, particularly for low-frequency variants in nearly 20 genes in the MHC region. This suggests that many of the genes providing suggestive or significant evidence of association in the MHC region likely have a mixture of deleterious, neutral, and protective variants within each gene and may explain why the burden test performed so poorly for genes in the MHC region. Conversely, it may be that the variants in the genes *TAGAP*, *RSPH3*, and *PAPOLG* tend to be unidirectional in effect and may explain why the

burden test was able to detect suggestive evidence of an association while the C-alpha test did not detect evidence of association at these genes. It is also worth noting how many low-frequency and rare variants were discovered at some genes, particularly, *LMOD3* (over 80% of the variants at this gene are either low-frequency or rare), *ABCF1* (80% of the variants at this gene are rare), and *MRPS18B* (nearly 90% of the variants at this gene have MAF < 0.01). Many of these low-frequency and rare variants are in intronic regions that would not have been detected by earlier studies such as Hunt et al.[75] but may play a key regulatory role and may help explain the heritability of celiac disease. However, the MAF for each variant from the sample in this study may be biased by the small sample size.

The aim of this analysis was to test the hypothesis that rare and low-frequency variants in previously identified candidate genes are associated with celiac disease using targeted resequencing data from 12 non-MHC regions as well as the MHC region. The results provide evidence for rare and low-frequency variants associated with celiac disease in 26 genes, three of which have previously been identified in previous studies. Further investigation will be required to validate the novel associations with larger studies, preferably re-sequenced or through imputation of regions with suggestive evidence of association into a larger previous GWAS.

Chapter 6: Analysis of Imputed Low-frequency and Rare Variants

As it is not always possible to directly genotype or sequence low-frequency and rare variants because of time and cost constraints, imputation has emerged as a very useful and cost-effective tool for human genetics researchers studying complex traits and diseases. While imputation of common (MAF $\geq$ 0.05) SNPs is now routinely done to impute both untyped and sporadically missing genotypes with very high accuracy, imputation of low-frequency and rare variants is still not a common practice because until recently, both the computational cost of running the algorithms for imputation and the lack of large, high-coverage reference panels to ensure accuracy for low-frequency and rare variants have been the main hurdles. Now, with computationally efficient algorithms, access to high-performance compute clusters at low cost, and reference panels that take advantage of thousands of haplotypes that cover the entire genome, imputation of low-frequency and rare variants should become a common practice. This study imputes low-frequency and rare variants to perform gene-based and single-marker association testing of celiac disease cases and controls from a previous celiac disease GWAS dataset.

6.1. Introduction

Recently, there has been a shift away from analysis of common variants to the analysis of low-frequency and rare variants in studies to determine the genetic susceptibility of many common, non-Mendelian diseases. With low-frequency and rare variants, a key issue that has come up is accurate genotyping as the allele frequencies fall because most of the microarray-based genotyping panels typically do not capture variants with a MAF < 0.05 with high accuracy. While denser microarrays have been developed that capture greater than 2 million variants and allow for detection of some low-frequency variants the gold standard for detecting low-frequency and rare variants remains sequencing and next-generation sequencing (NGS) although resequencing of large samples remains cost prohibitive at the moment[207]. The next best alternative is the imputation of missing

genotypes that were not directly genotyped on a previous GWAS microarray by using a dense reference panel. The imputation of common variants is now highly accurate with very high posterior probabilities for imputed genotypes. Imputation of low-frequency and rare variants with high accuracy remains a challenge[208,209] but methods and strategies have been developed[210,211] to increase the number of low-frequency and rare variants that are accurately imputed. Figure 6.1 provides a short overview of how a researcher may go about performing imputation and association testing of low-frequency and rare variants under current best-practices guidelines.

In the first step, thorough quality-control (QC) of the GWAS data for all samples with missing genotypes must be performed if the QC was not previously performed. Next, a reference panel must be selected to impute from. The reference panel may be from a previous GWAS, a large-scale sequencing project such as the 1000 Genomes Project, or a combination of reference panels. The third step involves phasing the haplotypes of the GWAS samples while the fourth step is the actual imputation of missing genotypes. While the third and fourth steps may be performed simultaneously in one software package, current best-practices guidelines suggest splitting the two tasks because of the considerable time-saved from phasing the haplotypes in a separate software packaged optimized for phasing. Finally, association analysis—single-variant testing or gene-based testing—may be performed on the imputed genotypes. Mägi et al.[201] successfully demonstrated that this approach could be used to impute rare variants (MAF < 0.01) into the WTCCC dataset and detect a genome-wide significant association of rare variants with coronary artery disease.

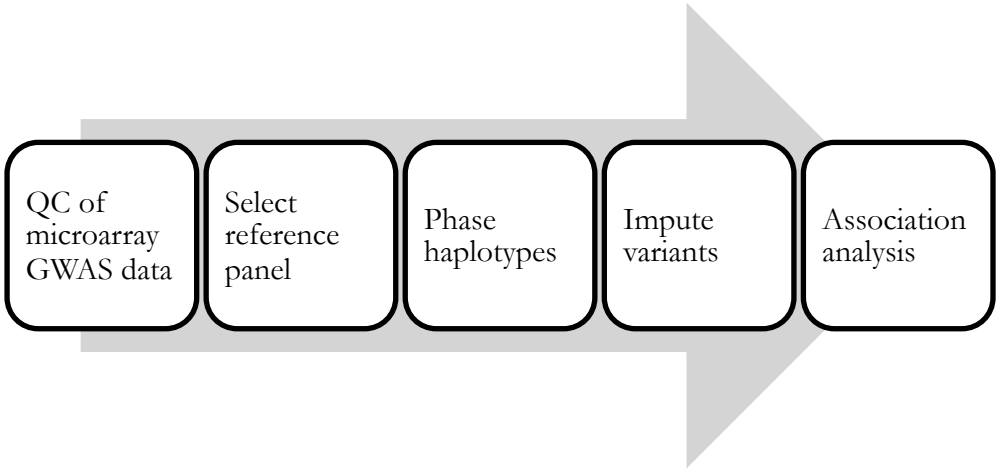


Figure 6.1. A simple flowchart of the steps to impute low-frequency and rare variants for association analysis.

Imputation algorithms

Multiple imputation (MI) is the general approach underlying modern genotype imputation. It is an easily extensible method that overcomes problems inherent in unprincipled, deterministic methods of imputation such as carrying last measured values forward or single random imputation that yield standard errors that are much too small and biases downstream analyses[212–215]. MI is independent of the analysis model and has been implemented in several software packages in diverse disciplines over the last few decades. In principle, MI works by sampling multiple independent sets of the missing data from the posterior predictive distribution of the missing data given the observed data that is typically done using computational algorithms such as Markov Chain Monte Carlo (MCMC) rather than analytical methods. While the true values of the missing data is never known with full certainty, MI does fully account for any uncertainty in the missing values. Another way of viewing MI is that it is a general approach for estimating missing-data uncertainty.

The IMPUTE v1[216] method was developed from a hidden Markov model (HMM) originally developed for simulating coalescent trees[217,218] and for linkage disequilibrium (LD) modeling. The HMM for each sample i's genotypes G_i is

$$Pr(G_i|H,\theta,\rho) = \sum_z Pr(G_i|Z,\theta), Pr(Z|H,\rho).$$

Here, Z is the vector of haplotype pairs from the reference panel, θ is the estimated mutation parameter, and ρ is the estimated fine-scale recombination rate. The $Pr\ (Z|H,\rho)$ term is the model for how the haplotype pairs change and is determined by a Markov chain that switches back and forth between states conditional upon the recombination map, ρ, of the genome. The conditional probability of genotype i, $P(G_i|Z,\theta)$, allows the observed genotype to be conditioned on the haplotype pairs that have been copied and have a mutation rate that is controlled by the parameter θ. The estimates for the fine-scale recombination rate are obtained from the IMPUTE 1 homepage. One must also define the effective population size, N_e. Finally, the exact marginal probability distribution for the missing genotype conditional on the observed genotype data in the vector, G_i, is obtained using the forward-backward algorithm[219].

IMPUTE v2[220], divides SNPs into two sets, T and U. T is the set of SNPs that are typed in the study sample and the reference panel whereas U is the set of SNPs that are untyped in the study sample but is genotyped in the reference panel. IMPUTE v2 uses the IMPUTE v1 HMM to estimate the haplotypes in T and proceeds to impute the alleles for SNPs in U by conditioning on the estimated haplotypes in T. The phasing of SNPs is performed by a Markov chain Monte Carlo (MCMC) algorithm in an alternating fashion where phasing and haploid imputation for a subset of SNPs. IMPUTE v2 can use either the haplotypes from the 1000 Genomes Project[221] or the haplotype sets from the HapMap3[153] as reference. More details of the IMPUTE v2 method will be covered below.

The MACH algorithm[222] implements an HMM model that is like the one used by IMPUTE. As it was designed to phase haplotypes, it is logically extended to perform imputation of missing genotypes as well. MACH iteratively updates the phase of each person's genotype conditioned on the haplotype of the other individuals in the study dataset. The model is,

$$P(G_i|D_{-i},\theta,\eta) = \Sigma P(G_i|Z,\eta)P(Z|D_{-i},\theta),$$

where D_{-i} is the set of all the estimated haplotypes leaving out the haplotype of individual i, Z is the unknown states of the HMM, η is a parameter that determines how much the copied haplotypes and G_i are alike, and θ is the parameter that controls the transitions between the hidden states. At each iteration of the algorithm, η and θ are updated based on changes in Z and the concordance rate between observed genotypes and the unknown genotypes that are in the hidden states. IMPUTE1 differs from MACH in that MACH does not use fixed estimates of mutation rates or recombination rates. MACH estimates the mutation rates and recombination rates based on the sample data and performs the genotype imputation via maximum-likelihood. While this gives MACH some more flexibility to adapt to the data as it is analyzed, there is a cost to this approach in terms of cost of imputation accuracy[210].

The BEAGLE algorithm[223] iteratively fits a model to the current set of estimated haplotypes and then resamples newly estimated haplotypes for each individual. The final missing genotypes are imputed from the model that has been fitted at the last iteration of the algorithm. Unlike IMPUTE, there are no parameters (i.e. H, Z, ρ, θ) that need to be estimated. BEAGLE works in two steps: the first involves creating a tree of haplotypes that bifurcates from left to right across all haplotypes in the sample dataset where each edge is weighted by the count of haplotypes that go along that edge. The second step involves the parsimonious pruning of the tree produced in the first step. The number of edges in any particular region is determined by the LD in that region and it is this property that the method works off; the model can be adapted to the haplotype diversity in any given dataset.

IMPUTE v2 has been shown to be the most accurate though it is only marginally more accurate than other methods[207,210,224]. While it has been noted[220,223] that the HMM models used by both IMPUTE v1 and MACH scale non-linearly as the number of haplotypes in the reference panel

increases, the adaptive haplotype selection approach in IMPUTE v2 scales linearly with the number of haplotypes in the panel and overcomes this problem. As a result IMPUTE v2 is much faster than BEAGLE or MACH.

IMPUTE v2 in detail

To estimate the posterior probabilities of the missing genotypes (set U), IMPUTE v2[220] switches between two steps, the first to phase the unphased SNPs and the second to impute missing genotypes by conditioning on the haplotype estimates from the phasing step. Let $H_R^{T,U}$ be the set of the reference haplotypes from both T and U, H_R^T the set of reference haplotypes that are only in set T, and H_S^T the set of unknown study haplotypes in set T. Given N_S in the study sample, then the haplotypes are represented, $H_S^T = \{H_{S,1}^T, \ldots, H_{S,N_S}^T\}$ where $H_{S,i}^T$ is the haplotype for each sample *i*. In the first step, initial estimates for H_S^T are made by sampling from the conditional distribution $\Pr(H_{S,i}^T | G_{S,i}^T, H_{S,(-i)}^T, H_R^T, \rho)$. After these initial haplotype guesses are made, Markov chain Monte Carlo (MCMC) iterations are made to update each sample *i* in two steps.

To be more specific, the first step in the MCMC iterations involves re-sampling $H_{S,i}^T$ for each sample *i* in *T* from the conditional distribution $Pr(H_{S,i}^T | G_{S,i}^T, H_{S,(-i)}^T, H_R^T, \rho)$, where $G_{S,i}^T$ is person *i*'s genotypes at the SNPs in *T*, $H_{S,(-i)}^T$ is the current haplotypes at SNPs for all individuals except person *i*, H_R^T represents the reference panel haplotypes at SNPs in *T*, and ρ is the population-scaled recombination map. In other words, this first step phases a sample *i*'s observed genotype by sampling from the conditional distribution, $Pr(H_{S,i}^T | G_{S,i}^T, H_{S,(-i)}^T, H_R^T, \rho)$, which is specified by a hidden Markov model (HMM)[216]. This HMM uses fixed recombination rates[225] for transition probabilities and a fixed mutation rate for the emission probabilities that is estimated from population genetics theory[216].

In the second step, new alleles for the SNPs in set U, conditional on $H_{S,i}^{T}, H_{R}^{T,U}$, and ρ are imputed for each of the two haploids for each sample i where the $H_{S,i}^{T}$ were sampled in step 1. Only the reference haplotypes $H_{R}^{T,U}$ are included in the state space of the HMM of step 2. For the actual imputation, the forward-backward algorithm is implemented in the i[th] haplotype in $H_{S,i}^{T}$. The posterior probabilities of the missing alleles are then estimated analytically. Allelic posterior probabilities can then be converted to genotypic posterior probabilities for each sample i under the assumption of Hardy-Weinberg Equilibrium. The missing genotype posterior probabilities are then summed across iterations, with a final posterior probability calculated by normalizing the summed posterior probabilities after all MCMC iterations are completed.

The computational burden of step one grows non-linearly ($O(N^2)$)as haplotypes are added and linearly (O(N)) as SNPs are added. To deal with the non-linear computational burden of phasing updates as haplotypes are added, the authors of IMPUTE v2 use an approximation of the conditional distribution that conditions the phasing updates on a subset of all haplotypes (rather than all haplotypes) in the sample space. To select the haplotypes for this subset, the algorithm first calculates the Hamming distance for each sample *i*'s best guess initial haplotype. The k haplotypes that have the smallest Hamming distance are then used by the HMM to sample new haplotypes for each sample. Recently, a more efficient phasing algorithm, SHAPEIT[226], that also takes advantage of multi-threaded processors (a given in contemporary computers, especially in high-performance compute clusters), has become part of the "best practices" pipeline by Howie et al. (http://mathgen.stats.ox.ac.uk/impute/impute_v2.2.2.html#best_practices_for_imputation) to "pre-phase" the study sample genotypes so that IMPUTE v2 only needs to perform imputation and not phasing as well. While there is a marginal tradeoff in terms of imputation accuracy with pre-phasing, the benefit tends to outweigh the cost, especially for datasets that contain tens of thousands of individual haplotypes to phase; whereas the phasing process using IMPUTE v2 alone may take

several days per chromosome, SHAPEIT can perform the same task in a day or less on a multi-core server. Another benefit to pre-phasing is that once a study sample chromosome has been phased, imputation of various stretches of the chromosome can be undertaken without having to perform the phasing each time.

Choice of reference panel

To optimally impute low-frequency and rare variants with high accuracy (i.e. variants with a high information metric or r^2), it is crucial to select the right reference panel. Before the advent of the 1000 Genomes Project[221], the standard for imputation reference panels was the latest HapMap dataset, the phase 3 reference panel[153]. The HapMap phase 3 (HMP3) data consists of genotype data from microarrays and sequencing: 1.6 million common variants from 1,184 samples as well as the sequence data for a subset of 692 of the 1,184 samples that comprises ten 100-kb regions. As the International HapMap Project was designed to catalog common variation across the genome for samples from African, Asian, European, and admixed populations from North and South America, the HMP3 data provides good coverage as a reference panel for common variants but provides sparse coverage for low-frequency and rare variants because for variants with MAF < 0.05 are less shared among related populations[227].

To surmount the limitation posed by imputing from HMP3 data, 1000 Genomes Project[221] (1KGP) data may be used as the reference panel instead. Use of the 1KGP data as a reference panel is preferable over HMP3, especially for imputing low-frequency and rare variants because the 1KGP was designed to identify the majority (>95%) of variants that have a MAF ≥ 0.01 across the entire genome and MAF<0.01 within specific gene regions. The most recent published update to the 1KGP is the Phase 1 dataset[154] that includes data from 1,092 genomes (from European, African, and Asian populations) with the final phase to include data from 2,500 genomes. Studies by Jostins et al.[228] and Sung et al.[209], have shown a four-fold increase in successfully imputed low-frequency

variants (MAF < 0.05) and an eight-fold increase in successfully imputed rare variants (MAF ≤ 0.01) by employing the pilot 1KGP data (283 genomes) rather than the HMP3 data.

Post-imputation

After imputation of missing genotypes has been carried out, but before association testing of variants, an additional quality control step to assess the imputed genotypes must be taken to filter out poorly imputed variants when a set of directly genotyped variants is not available to compare the imputed genotypes to[13]. While each imputation method produces its own imputation quality metric, all metrics have a range between 0 and 1, where 0 indicates that there is no certainty about the imputed genotype and a value of 1 indicates perfect certainty about the imputed genotype by essentially comparing the observed variance of the estimated genotypes to the expected variance. It has been shown in the literature that all of the imputation quality metrics are highly correlated with each other.[210]

IMPUTE (v1 and v2) calculates a metric that is based on comparing the unknown population allele frequency θ_j to the estimated allele frequency $\widehat{\theta_j}$. To calculate the information measure I_A, the score, $U(\theta)$, and information, $I(\theta_j)$, of the full data likelihood, $L(\theta_j)$ must be derived where

$$L(\theta_j) = \prod_{i=1}^{N} \theta_j^{G_{ij}} (1-\theta_j)^{2-G_{ij}},$$

$$U(\theta_j) = \frac{d\, log\, L(\theta_j)}{d\theta_j} = \frac{X-2N\theta_j}{\theta_j(1-\theta_j)},$$

$$I(\theta_j) = \frac{-d^2\, log\, L(\theta_j)}{d\theta_j^2} = \frac{X}{\theta_j^2} + \frac{2N-X}{(1-\theta_j)^2}.$$

Finally, the measure I_A may be calculated as follows,

$$I_A = \frac{(E[I(\widehat{\theta})] - Var[U(\widehat{\theta})])}{E[I(\widehat{\theta})]}.$$

Here $\hat{\theta} = \frac{\Sigma_{i=1}^{N} e_{ij}}{2N}$, G_{ij} is the genotype for individual *i* at SNP *j*, and $X = \Sigma_{i=1}^{N} G_{ij}$. I_A has an upper bound of 1 and equals 0 when the mean variance of the imputed genotypes equals the expected variance if the alleles were sampled at the estimated allele frequency (i.e. total uncertainty for the given imputed genotype).

The information measure produced by MACH is the $\hat{r}^2$ measure[222] and it is the ratio of the empirically observed variance of allele dosage at the *j*th SNP against the expected variance given Hardy-Weinberg equilibrium,

$$\hat{r}_j^2 = \begin{cases} \frac{\frac{\Sigma_{i=1}^{N} e_{ij}^2}{N} - \left(\frac{\Sigma_{i=1}^{N} e_{ij}}{N}\right)^2}{2\hat{\theta}(1-\hat{\theta})} \; when\; \hat{\theta} \in (0,1) \\ 1 \; when\; \hat{\theta} = 0, \hat{\theta} = 1 \end{cases}.$$

Here e_{ij} is the expected allele dosage for a genotype at SNP *j* for individual *i*. While this ratio is nearly unity when there is high certainty for a genotype, it can actually go above one[210].

BEAGLE's information metric, R^2, is an approximation of the correlation between the best estimate of the genotype and the actual genotype[223]. For the *j*th SNP this metric is defined,

$$R_j^2 = \frac{\left[\Sigma_i z_{ij} e_{ij} - \left(\frac{1}{N}\right)\left(\Sigma_i z_{ij} \Sigma_i e_{ij}\right)\right]^2}{\left[\Sigma_i f_{ij} - \left(\frac{1}{N}\right)\left(\Sigma_i e_{ij}\right)^2\right]\left[\Sigma_i z_{ij}^2 - \left(\frac{1}{N}\right)\left(\Sigma_i z_{ij}\right)^2\right]},$$

where the z_{ij} term is the imputed genotype in the *i*th individual at the *j*th SNP with the highest likelihood.

Association testing methods

With univariate tests it is often difficult to detect associations with low-frequency and rare variants (i.e. chi-square test or logistic regression). For single variant tests, packages such as SNPTEST account for the uncertainty in the imputed genotypes[208]. Alternatively, one may also filter out the low-quality imputed genotypes by applying some arbitrary cutoff threshold for whichever information metric is being used to assess the imputation quality at a given SNP. To gain more

power over univariate tests, low-frequency and rare variants may be combined across a gene (or some other unit, though it is usually gene-level) and this aggregated or collapsed variable is then tested for an accumulation of low-frequency or rare variants (discussed previously in chapter 4).

The present study attempted to identify imputed rare and low-frequency variants associated with celiac disease using a large dataset previously used in a GWAS[73]. Results from chapter 4 were used to determine which genes to impute in the large dataset. Both single-marker and gene-based tests were implemented to test for association. This study did not attempt to impute all rare and low-frequency variants genome-wide.

6.2. Methods

GWAS dataset

The study genotypes employed for imputing forward missing low-frequency and rare variants come from the Trynka et al.[73] dense genotyping dataset that was previously employed in the meta-analysis study in chapter 3. Briefly, this dataset is comprised of 12,041 cases and 12,228 controls (n = 24,269) from seven distinct sample collections and contains 139,553 SNPs. This dataset has already been through stringent quality control to remove low-quality variants and low-quality samples. As this dataset has genomic coordinates from the b36/hg18 genome assembly, the liftOver[229] program was used to map variant positions to those of the b37/hg19 genome assembly, the assembly that is used for the 1KGP Phase 1 reference panel. GTOOL[220], a program that accompanies the IMPUTE v2 package, was used to format the GWAS dataset for input into IMPUTE v2. Following best practice guidelines for efficient and accurate imputation using IMPUTE v2, the SHAPEIT2[226,230] package was implemented to pre-phase the haplotypes in the Trynka et. al[73] dataset. Default algorithm and model parameters were used: 7 burn-in iterations, 8 iterations for pruning the genotype graphs by the transition probabilities, 20 main iterations to estimate the final haplotype, 100 conditioning states per

SNP, a window size of 2 Mb, and an effective population size of 11,418 (estimated from HapMap populations for CEU).

Imputation

The IMPUTE v2 package was used to impute forward the missing low-frequency and rare variants in the Trynka et al.[73] dataset. Best practice guidelines provided by the authors of IMPUTE v2 were followed to set algorithm parameters. The effective population size, N_e, was set to 20,000, threshold for calling genotypes from the study sample was set at 0.9, the number of MCMC iterations set at 30 (with 10 burn-in iterations), and the number of reference haplotypes to use as templates when imputing missing genotypes was set at 1000. The following genomic intervals were imputed: chromosome 2: 60,983,365-61,029,221 (*PAPOLG*); chromosome 3: 69,134,090-69,172,183 (*ARL6IP5* and *LMOD3*); chromosome 4: 123,073,488-123,283,914 (*KIAA1109*) chromosome 6: 30,539,153-32,374,905 (part of the classical MHC region) and 159,393,903-159,466,184 (*RSPH3* and *TAGAP*). These chromosomal regions were chosen for imputation because they yielded suggestive or significant evidence of association under gene-based association testing in chapter 4.

Post-imputation filtering

Filtering of imputed variants by the info score, I_A, and subsequent conversion to the PED or VCF formats to be read into association testing packages was performed by GTOOL/QCTOOL. While there is no universally accepted cut-off value for a well-imputed variant, the literature suggests that an I_A threshold greater than 0.3 is a widely used threshold for filtering out poorly-imputed variants[13,201,222,231,232]. Li et al.[222] demonstrated that 70% of poorly imputed variants would be filtered out while less than 1% of well imputed variants would be filtered out with a threshold greater 0.3. For this study, variants were excluded from the final analysis dataset if $I_A < 0.4$.

Power analysis

A power analysis was carried out using the CaTS package[233] using the following fixed parameters: 250 cases and 250 controls to represent the resequenced dataset, 10000 cases and 10000 controls to represent the imputed dataset (maximum allowed by the software; 12,041 cases and 12,228 controls were included in this study), significance level of $p = 5x10^{-8}$, assumed prevalence of disease set at 0.01. Power was plotted as a function of the MAF (range from 0 to 0.3) for various genotype relative risk levels (range from 1.2 to 3) under an additive model.

Gene-based association testing

Analysis of low-frequency and rare variants was performed using the two collapsing type methods that were previously implemented in chapter 4, the fixed threshold burden test of Morris et al.[187] and the C-alpha test of Neale et al.[188]. As described before, p-values were asymptotically obtained under the fixed threshold burden test and empirically under the C-alpha test. Under the C-alpha test, 10000 permutations were performed for each gene. Under either test, variants were grouped by gene and fine-scale QC was performed as previously described. A MAF threshold of 0.01 was set for rare variant testing while a threshold up to 0.04 was set for low-frequency variant testing as rare variants are defined as having MAF < 0.01 while low-frequency variants have MAF < 0.05. To adjust for sample collection membership, an indicator variable for each of the seven collections was included in both the burden and C-alpha tests. As approximately 60 genes were tested, a p-value $\leq 8x10^{-4}$ (after Bonferroni correction) was a statistically significant finding; any finding with a p-value greater than or equal to $1x10^{-3}$ was considered statistically suggestive evidence.

Single-marker association testing

A Cochran-Mantel-Haenszel based fixed-effects meta-analysis method implemented in PLINK[234] that incorporates a sample collection membership indicator variable to account for ethnic differences in a logistic regression framework was utilized. An established, conservative p-value ≤

$5x10^{-8}$ was considered genome-wide significant (GWS) evidence while a p-value > $5x10^{-8}$ and ≤ $1x10^{-6}$ was considered borderline GWS evidence[235].

6.3. Results

Imputation

Imputation using the 1KGP Phase 1 reference panel[154] yielded 25,104 successfully imputed variants across the 6 genes (on chromosomes 2, 3, and 6) and much of the MHC region. Of these imputed variants, 18,792 variants (approximately 75%) had an IMPUTE v2 info metric, $I_A \geq 0.4$ (table 6.1). High-quality imputation was not uniform across genes as evidenced by the difference between the percentage of variants with $I_A \geq 0.4$, where for instance, *LMOD3* only had 1.9% of variants well-imputed while *RSPH3* had 63.4% of variants well-imputed. Of these 18,792 well-imputed variants, 7,117 (approximately 38%) had MAF ≤ 0.04 and 4,321 (approximately 23%) had MAF < 0.01 (table 6.2).

Table 6.1. Total number of variants successfully imputed by gene/region and the proportion that are well-imputed.

Gene/Genomic Region	Total Number Imputed	Imputed with $I_A \geq 0.4$ (%)
ARL6IP5	290	38 (13.1%)
LMOD3	258	5 (1.9%)
PAPOLG	502	148 (29.5%)
KIAA1109	4911	2328 (47.4%)
RSPH3	421	267 (63.4%)
TAGAP	148	73 (49.3%)
MHC region	23,485	18,261 (77.8%)

Table 6.2. Proportion of well-imputed variants that are low-frequency or rare.

Gene/Genomic Region	MAF ≤ 0.04	MAF < 0.01
ARL6IP5	22/38	21/38
LMOD3	1/5	1/5
PAPOLG	50/148	29/148
KIAA1109	1468/2328	1207/2328

RSPH3	164/267	49/267
TAGAP	29/73	18/73
MHC region	6,851/18,261	4,203/18,261

Gene-based association test results

Under the rare burden test, two genes provided significant evidence of association while another two genes provided suggestive evidence of association (table 6.3); only one of the genes with either significant or suggestive evidence was not from the MHC region (*RSPH3*). When the MAF threshold for the burden test was increased to 0.04, p_{POU5F1} decreased and two other genes yielded significant evidence of association (table 6.4). All of the genes that yielded significant or suggestive evidence of association for the low-frequency burden test were from the MHC region. When only the non-synonymous (NS) variants are included in the rare burden test, one gene, *AGER* ($p = 1.24\text{x}10^{-4}$) had significant evidence. Under a low-frequency burden test of NS variants, one gene, *POU5F1* ($p = 4.52\text{x}10^{-5}$) had significant evidence of association while another two genes, *LTB* ($p = 8.60\text{x}10^{-4}$) and *RSPH3* ($p = 3.43\text{x}10^{-3}$) yielded suggestive evidence.

The rare C-alpha test yielded 35 genes with significant evidence of association with another 3 genes providing suggestive evidence of association (table 6.5). All of these genes were located within the MHC region. The low-frequency C-alpha test (table 6.6) provided significant evidence of association for 46 genes and suggestive evidence for one more gene, all located within the MHC region. When the rare C-alpha test was restricted to NS variants only, 12 genes provided significant evidence of association while another 2 had suggestive evidence (table 6.7). The low-frequency C-alpha test of only NS variants yielded 18 genes with significant evidence of association and 3 genes with suggestive evidence (table 6.8). Across the rare variants that were significant or suggestive, the average of the proportion of imputed variants for a given gene that are rare was approximately 40%

while for low-frequency variants this average is approximately 55%; the proportions are approximately 56% and 69%, respectively, when only NS variants are considered.

Table 6.3. Results of burden test with MAF < 0.01.

Gene	Variants tested/total variants (%)	P_{Burden}
POU5F1	26/96 (27%)	$2.45x10^{-5}$
LTB	2/3 (67%)	$1.92x10^{-3}$
PSORS1C3	12/72 (17%)	$2.25x10^{-4}$
RSPH3	57/214 (27%)	$2.53x10^{-3}$

Table 6.4. Results of burden test with MAF ≤ 0.04.

Gene	Variants tested/total variants (%)	P_{Burden}
C6orf47	9/12 (75%)	$1.58x10^{-6}$
POU5F1	36/96 (38%)	$7.45x10^{-6}$
ATP6V1G2	7/14 (50%)	$3.03x10^{-5}$
CDSN	29/103 (28%)	$1.11x10^{-3}$

Table 6.5. Results of C-alpha test with MAF < 0.01.

Gene	Variants tested/total variants (%)	$P_{C\text{-}alpha}$
AGER	9/20 (45%)	$< 1.00x10^{-4}$
AGPAT1	16/33 (47%)	$< 1.00x10^{-4}$
APOM	13/26 (50%)	$< 1.00x10^{-4}$
C6orf15	4/22 (18%)	$< 1.00x10^{-4}$
ATF6B	31/54 (57%)	$< 1.00x10^{-4}$
C6orf47	8/12 (67%)	$< 1.00x10^{-4}$
CSNK2B	1/3 (33%)	$< 1.00x10^{-4}$
EGFL8	10/22 (45%)	$< 1.00x10^{-4}$
BAG6	38/79 (48%)	$< 1.00x10^{-4}$
GPANK1	21/47 (45%)	$< 1.00x10^{-4}$
HCG22	24/100 (24%)	$< 1.00x10^{-4}$
GPSM3	9/19 (47%)	$< 1.00x10^{-4}$
LOC100507547	2/6 (33%)	$< 1.00x10^{-4}$
HCG27	27/118 (23%)	$< 1.00x10^{-4}$
LTA	5/14 (36%)	$< 1.00x10^{-4}$
MCCD1	13/20 (65%)	$< 1.00x10^{-4}$
MICB	36/189 (19%)	$< 1.00x10^{-4}$
PBX2	10/25 (40%)	$< 1.00x10^{-4}$
PPT2	19/44 (43%)	$< 1.00x10^{-4}$
NOTCH4	56/182 (31%)	$< 1.00x10^{-4}$
PPT2-EGFL8	28/68 (40%)	$< 1.00x10^{-4}$

MUC22	94/534 (18%)	$< 1.00x10^{-4}$
PRRC2A	44/106 (42%)	$< 1.00x10^{-4}$
MICA	133/374 (36%)	$< 1.00x10^{-4}$
RNF5	3/13 (23%)	$< 1.00x10^{-4}$
PSORS1C3	12/72 (17%)	$< 1.00x10^{-4}$
SNORA38	1/2 (50%)	$< 1.00x10^{-4}$
TNF	6/11 (55%)	$< 1.00x10^{-4}$
TCF19	14/51 (27%)	$< 1.00x10^{-4}$
HCG23	74/132 (56%)	$< 1.00x10^{-4}$
PSORS1C1	119/407 (29%)	$< 1.00x10^{-4}$
TNXB	139/288 (48%)	$< 1.00x10^{-4}$
C6orf10	378/1042 (36%)	$< 1.00x10^{-4}$
BTNL2	262/398 (66%)	$< 1.00x10^{-4}$
CCHCR1	47/211 (22%)	$3.00x10^{-4}$
POU5F1	26/96 (27%)	$1.50x10^{-3}$
FKBPL	5/10 (50%)	$1.60x10^{-3}$
LTB	2/3 (67%)	$3.50x10^{-3}$

Table 6.6. Results of C-alpha test with MAF $\leq$ 0.04.

Gene	**Variants tested/total variants (%)**	$\mathbf{P_{C\text{-}alpha}}$
ATP6V1G2	7/14 (50%)	$< 1.00x10^{-4}$
AGPAT1	19/33 (58%)	$< 1.00x10^{-4}$
AGER	11/20 (55%)	$< 1.00x10^{-4}$
APOM	15/26 (58%)	$< 1.00x10^{-4}$
C6orf15	7/22 (32%)	$< 1.00x10^{-4}$
C6orf47	10/12 (83%)	$< 1.00x10^{-4}$
CSNK2B	2/3 (67%)	$< 1.00x10^{-4}$
CDSN	29/103 (28%)	$< 1.00x10^{-4}$
FKBPL	6/10 (60%)	$< 1.00x10^{-4}$
ATF6B	41/54 (76%)	$< 1.00x10^{-4}$
EGFL8	15/22 (68%)	$< 1.00x10^{-4}$
DDX39B	72/95 (76%)	$< 1.00x10^{-4}$
GPSM3	10/19 (53%)	$< 1.00x10^{-4}$
BAG6	50/79 (63%)	$< 1.00x10^{-4}$
HCG22	37/100 (37%)	$< 1.00x10^{-4}$
HCP5	17/35 (49%)	$< 1.00x10^{-4}$
GPANK1	32/47 (68%)	$< 1.00x10^{-4}$
LOC100507547	2/6 (33%)	$< 1.00x10^{-4}$
LST1	10/17 (59%)	$< 1.00x10^{-4}$
LTA	7/14 (50%)	$< 1.00x10^{-4}$
LTB	3/3 (100%)	$< 1.00x10^{-4}$
HCG27	40/118 (34%)	$< 1.00x10^{-4}$

MCCD1	16/20 (80%)	$< 1.00x10^{-4}$
MICB	71/189 (38%)	$< 1.00x10^{-4}$
NCR3	19/30 (63%)	$< 1.00x10^{-4}$
HCG23	121/132 (92%)	$< 1.00x10^{-4}$
CCHCR1	71/211 (34%)	$< 1.00x10^{-4}$
PPT2	24/44 (55%)	$< 1.00x10^{-4}$
PBX2	17/25 (68%)	$< 1.00x10^{-4}$
NOTCH4	81/182 (45%)	$< 1.00x10^{-4}$
PPT2-EGFL8	39/68 (57%)	$< 1.00x10^{-4}$
MUC22	175/534 (33%)	$< 1.00x10^{-4}$
PSORS1C2	15/31 (48%)	$< 1.00x10^{-4}$
RNF5	5/13 (38%)	$< 1.00x10^{-4}$
PSORS1C3	24/72 (33%)	$< 1.00x10^{-4}$
SNORA38	1/2 (50%)	$< 1.00x10^{-4}$
TCF19	19/51 (37%)	$< 1.00x10^{-4}$
TNF	8/11 (73%)	$< 1.00x10^{-4}$
PRRC2A	65/106 (61%)	$< 1.00x10^{-4}$
PSORS1C1	191/407 (47%)	$< 1.00x10^{-4}$
BTNL2	366/398 (92%)	$< 1.00x10^{-4}$
C6orf10	470/1042 (45%)	$< 1.00x10^{-4}$
MICA	219/374 (59%)	$< 1.00x10^{-4}$
TNXB	179/288 (62%)	$< 1.00x10^{-4}$
POU5F1	36/96 (38%)	$< 1.00x10^{-4}$
ATP6V1G2-DDX39B	81/120 (68%)	$5.60x10^{-3}$

Table 6.7. Results of C-alpha test with MAF < 0.01 including only non-synonymous variants.

Gene	**Variants tested/total variants (%)**	$\mathbf{P_{C\text{-}alpha}}$
AGER	2/3 (67%)	$< 1.00x10^{-4}$
C6orf15	2/12 (17%)	$< 1.00x10^{-4}$
CDSN	5/10 (50%)	$< 1.00x10^{-4}$
ATF6B	5/5 (100%)	$< 1.00x10^{-4}$
BAG6	2/3 (67%)	$< 1.00x10^{-4}$
PSORS1C1	12/25 (48%)	$< 1.00x10^{-4}$
BTNL2	19/31 (61%)	$< 1.00x10^{-4}$
C6orf10	11/22 (50%)	$< 1.00x10^{-4}$
TNXB	16/38 (42%)	$< 1.00x10^{-4}$
NOTCH4	7/14 (50%)	$2.00x10^{-4}$
CCHCR1	9/25 (36%)	$5.00x10^{-4}$
LTB	2/2 (100%)	$1.80x10^{-3}$
MUC22	14/44 (32%)	$2.40x10^{-3}$
FKBPL	2/3 (67%)	$2.60x10^{-3}$

Table 6.8. Results of C-alpha test with MAF ≤ 0.04 including only non-synonymous variants.

Gene	Variants tested/total variants (%)	$P_{C\text{-}alpha}$
AGER	3/3 (100%)	$< 1.00x10^{-4}$
C6orf15	5/12 (42%)	$< 1.00x10^{-4}$
CDSN	5/10 (50%)	$< 1.00x10^{-4}$
ATF6B	5/5 (100%)	$< 1.00x10^{-4}$
BAG6	2/3 (67%)	$< 1.00x10^{-4}$
PSORS1C1	15/25 (60%)	$< 1.00x10^{-4}$
BTNL2	29/31 (94%)	$< 1.00x10^{-4}$
C6orf10	12/22 (55%)	$< 1.00x10^{-4}$
TNXB	24/38 (63%)	$< 1.00x10^{-4}$
GPANK1	4/5 (80%)	$< 1.00x10^{-4}$
MCCD1	3/4 (75%)	$< 1.00x10^{-4}$
MICB	3/8 (38%)	$< 1.00x10^{-4}$
MICA	14/29 (48%)	$< 1.00x10^{-4}$
PSORS1C2	7/12 (58%)	$< 1.00x10^{-4}$
PRRC2A	13/19 (68%)	$< 1.00x10^{-4}$
NOTCH4	10/14 (71%)	$< 1.00x10^{-4}$
MUC22	25/44 (57%)	$< 1.00x10^{-4}$
CCHCR1	12/25 (48%)	$< 1.00x10^{-4}$
LTB	2/2 (100%)	$1.50x10^{-3}$
NCR3	5/5 (100%)	$1.90x10^{-3}$
FKBPL	2/3 (67%)	$1.50x10^{-3}$

Single-marker association test results

GWS evidence ($p \leq 5x10^{-8}$) of association with celiac disease with low-frequency or rare variants was found in 23 SNPs across 7 genes, all non-HLA genes within the classical MHC region; there was no GWS evidence of association with celiac disease in any of the non-MHC (table 6.9). Another two SNPs, both in the gene *ABCF1* had borderline GWS evidence of association. Most of the GWS SNPs are low-frequency (MAF ≤0.04) with just four SNPs, one in gene *C6orf10* and three in *BTNL2*, having MAF ≤ 0.01. All four of these rare variants were associated with a reduced risk of developing celiac disease; most of the GWS variants were associated with a reduced risk of

developing celiac disease. The four of five GWS (and two suggestive SNPs) SNPs associated with celiac disease had an OR that indicated a modest increase in risk with ORs ranging between approximately 2 and 3. There was one standout SNP, rs17201553 of *TNXB*, with an OR of 4.40. It is an intronic variant, as are most of the GWS SNPs, with only four exonic variants (all missense). Only one of the SNPs with suggestive or significant evidence of association, rs1264446, was also tested in a gene-based test using resequencing data (chapter 5). As expected from the power analysis (figure 6.2), there were no suggestive or significant low-frequency or rare SNPs with an OR < ~2 as even with ~24K samples (imputed dataset), this study is underpowered to detect GWS signals using a single-marker test. However, the power analysis had also revealed that at an OR of 2 or 3, the imputed dataset would increase the power to detect an association with rare variants from 0% to 100% relative to the resequenced dataset. When considering an OR of 3, the imputed dataset would increase the power to detect an association with a low-frequency variant from 15% to 100%.

Table 6.9. Fixed-effects (Cochran-Mantel-Haenszel) single-marker association results for low-frequency and rare variant association with celiac disease.

Gene	SNP	Alleles	Position	MAF	P_{CMH}	OR (95% CI)	Type
ABCF1	rs1264446	A/G	6:30546395	0.013	1.24E-07	2.0 (1.53 – 2.60)	Intronic
	rs1264450	C/T	6:30541761	0.013	3.12E-07	1.96 (1.50 – 2.55)	Intronic
HCP5	rs2395029	G/T	6:31539759	0.024	9.49E-09	0.50 (0.40 – 0.64)	Missense
AIF1	rs28732150	A/G	6:31691203	0.039	2.59E-22	0.42 (0.35 – 0.50)	Intronic
BAT2	rs2280801	A/G	6:31700043	0.039	1.81E-23	0.41 (0.34 – 0.49)	Missense
TNXB	rs41270450	C/G	6:32125151	0.026	2.53E-16	0.41 (0.33 – 0.51)	Missense
	rs41316748	C/T	6:32127490	0.024	1.40E-14	0.38 (0.29 – 0.49)	Intronic
	rs9267796	C/T	6:32131403	0.030	6.47E-26	2.47 (2.08 – 2.93)	Intronic
	rs28732167	A/G	6:32139882	0.032	1.53E-10	0.53 (0.44 – 0.64)	Intronic
	rs17201553	A/G	6:32148971	0.011	3.08E-42	4.40 (3.50 – 5.53)	Intronic
	rs9267799	C/T	6:32154922	0.030	1.19E-25	2.47 (2.08 – 2.94)	Missense
	rs17201560	C/T	6:32155246	0.029	3.93E-12	0.45 (0.36 – 0.57)	Intronic
	rs57740770	G/T	6:32175413	0.028	2.38E-26	2.56 (2.14 – 3.06)	Intronic
	rs28732173	A/G	6:32178988	0.026	9.04E-16	0.37 (0.29 – 0.47)	Intronic
PPT2	rs10947233	G/T	6:32232402	0.028	3.72E-10	0.32 (0.22 – 0.46)	Intronic
C6orf10	rs11751697	C/T	6:32374403	0.030	2.05E-23	0.40 (0.33 – 0.48)	Intronic
	rs3749967	C/T	6:32391822	0.030	2.05E-23	0.40 (0.33 – 0.48)	Intronic
	rs9268233	A/G	6:32397270	0.029	1.17E-38	3.21 (2.67 – 3.87)	Intronic
	rs28732193	C/T	6:32414900	0.025	1.26E-17	0.34 (0.26 – 0.44)	Intronic
	rs13196329	A/C	6:32433349	0.014	1.66E-09	0.49 (0.39 – 0.62)	Intronic

	rs2076535	A/G	6:32447489	0.004	1.90E-18	0.12 (0.06 – 0.21)	UTR-5'
BTNL2	rs13198563	A/G	6:32468893	0.009	9.40E-09	0.42 (0.31 – 0.57)	Intronic
	rs2076531	C/T	6:32471690	0.004	2.52E-20	0.11 (0.06 – 0.20)	Intronic
	rs2076528	G/T	6:32472172	0.004	8.68E-20	0.11 (0.06 – 0.20)	Intronic
	rs3763308	A/G	6:32482618	0.030	9.96E-11	0.57 (0.48 – 0.67)	Intronic

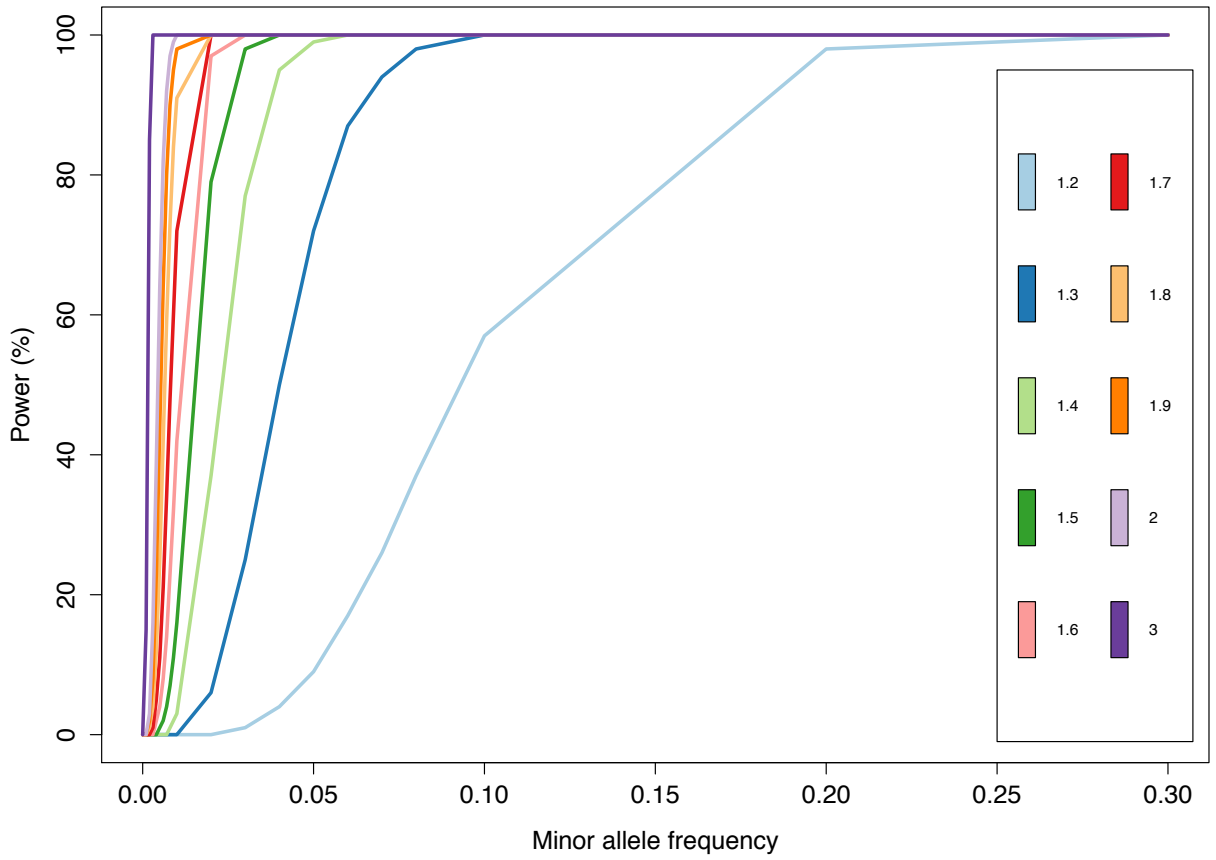


Figure 6.2. Power to detect an association as a function of the MAF for various genotype relative risks (color-coded).

6.4. Discussion

This study demonstrates that imputation of low-frequency and rare variants for association testing is not only feasible, but a desirable model, given that the cost to directly genotype or re-sequence samples to capture these low-frequency and rare variants remains high for the moment. Datasets

from previous GWASs with tens of thousands (some with hundreds of thousands) of samples are available and with high coverage reference panels like the 1KGP Phase 1 (and soon, the complete 2,500 sample completed 1KGP panel), it is now a matter of imputing forward the missing genotypes in the large-scale GWAS datasets to test refined hypotheses on the etiology of complex traits and diseases using the low-frequency and rare variants and generate preliminary data for future projects. While few studies have employed imputation of low-frequency and rare variants[201,207] for meta-analysis, there is active research in this area with new as-of-yet unpublished methods to perform imputation using population-scale resource panels such as the data from the UK10K Project (http://www.uk10k.org) which seeks to sequence the full genomes of 4,000 individuals and exome sequence another 6,000 individuals and separately to combine the multiple large-scale reference panels such as the 1KGP and UK10K resources and perform what is known as "meta-imputation"[236].

This study provides suggestive evidence that there are probably many low-frequency and rare variants, particularly in non-coding regions, that may increase understanding of the etiology of celiac disease as well as its heritability and have yet to be uncovered because previous GWAS of celiac disease have focused on common variants or variants within coding regions, such as the exome-sequencing based study conducted recently[75]. Under gene-based association tests of the imputed rare and low-frequency variants, one of the genes that yielded suggestive evidence of association with celiac disease in chapter 4, *RSPH3*, also provided suggestive evidence of association in the present study. Within the MHC region, nearly all of the genes that yielded either significant or suggestive evidence of association under the rare or low-frequency C-alpha test in chapter 4 were also significant when the imputed rare and low-frequency variants were tested, with the exception of genes *COL11A2* and *TRIM40*. In addition to the 20 MHC genes identified by the C-alpha test in

chapter 4, this present study identified an additional 26 genes that have either significant or suggestive evidence of association.

Interestingly, whereas the burden test yielded no significant or suggestive evidence of association for genes within the MHC region in chapter 4 with the resequenced data, the rare and low-frequency burden tests of the imputed variants provided evidence of six genes associated with celiac disease within the MHC region, one of which was identified when only the NS variants were included in the analysis, indicating that there may be rare or low-frequency variants within the MHC region that are unidirectional and that there may not have been sufficient depth of coverage in the resequencing study, inadequate sample size, or both. When only the imputed NS variants were included in the rare and low-frequency C-alpha tests, subsets of the genes found to have significant or suggestive evidence of association in the analysis of all imputed variants yielded evidence of association. No additional genes were discovered when only the NS variants were included in the C-alpha tests. These observations provide evidence that rare and low-frequency NS variants may play a significant role in the etiology and severity of celiac disease. As noted before in chapter 4, for many of the genes with evidence of association, a large proportion have a MAF ≤ 0.04, and a small sample size may not be biasing the MAFs observed as was possibly the case with the much smaller reseqeuncing study in chapter 4. However, it may be the case that some of the rare and low-frequency variants observed are specific to one of the seven sample collections that comprise the dataset.

The rare c-alpha test of the imputed variants yielded approximately 9 times more genes with significant evidence of association than a rare burden test while the low-frequency C-alpha test yielded approximately 12 times more genes with significant evidence of association than a low-frequency burden test. Given that the C-alpha test was designed to detect an association with disease in the presence of a mixture of protective, neutral, and deleterious variants, the results seem to

suggest that many of the imputed genes harbor rare and low-frequency variants that are not likely to be detected by a unidirectional test such as the burden test. However, while the rare and low-frequency C-alpha tests only detected associations at genes within the MHC region, the rare burden test was able to detect suggestive evidence of association at one non-MHC gene, *RSPH3*, which—as noted above—was the only gene to provide suggestive evidence of association in both chapter 4 with the resequenced variants and in the present study.

Of the single-marker test results, the most interesting signals were the very rare ($p = 0.004$) SNPs associated with celiac disease that are harbored in the non-coding regions of *C6orf10* and *BTNL2* as well as the highly significant and high effect size intronic SNPs from *TNXB* (rs17201553 with MAF of 0.011 and OR of 4.40 with $p = 3.08 \times 10^{-42}$) and *C6orf10* (rs9268233 with MAF of 0.029 and OR of 3.21 with $p = 1.17 \times 10^{-38}$). While the power analysis indicates that even this study is underpowered to detect rare variants with effect sizes < 2 with GWS, there is an indication that this study is well-powered to detect rare variants with effect sizes ≥ 2. The low-frequency SNPs from *HCP5* and *BAT2* that provided GWS evidence of association in the single-marker test along with the significant signals for the genes *GPSM3*, *HCP5*, and *PSORS1C1* under the rare and low-frequency C-alpha test of the imputed variants seem to strengthen the findings for common SNPs in a fine-mapping study of celiac disease conducted previously[74]. In the previous fine-mapping study, the common SNPs from *HCP5* and *BAT2* also were associated with lower risk of disease (OR of 0.57 and 0.60, respectively). While the C-alpha test implemented in chapter 4 failed to find significant evidence of association between rare variants and celiac disease except in *ABCF1*, both the single-marker test and the gene-based burden and C-alpha tests of the imputed rare variants have provided GWS evidence of association with disease.

As mentioned above, imputation performance may be increased in future studies by combining multiple large reference panels. The authors of IMPUTE also note that while pre-phasing

is preferred because the marginal loss in accuracy is outweighed by the time saved (and ostensibly making a large-scale imputation project feasible in a finite amount of time), it may be worthwhile to go back and re-run the imputation for regions with suggestive evidence (such as *ABCF1*) by re-running IMPUTE v2 with both the phasing and imputation steps (i.e. no pre-phasing via SHAPEIT). However, as IMPUTE v2 takes several orders of magnitude more time to perform phasing and imputation for each chromosome for 24K individuals, this author has found this approach infeasible in the limited timeframe of this dissertation project. It should be noted that the imputation performance (in terms of yielding well-imputed genotypes) for the genes/regions that were imputed were comparable to the performance seen by Mägi et al.[201]. While this study was focused on discovering imputed low-frequency and rare variants that are associated with celiac disease, a future study may focus on gene-based, genome-wide (i.e. not restricted to certain genes or regions) meta-analysis as implemented in an unpublished method such as RAREMETAL that was developed for the GSCAN project ("GWAS & Sequencing Consortium of Alcohol and Nicotine use"; http://gscan.sph.umich.edu). The RAREMETAL package was written to implement gene-based tests (burden based, variable threshold, SKAT) in an meta-analysis framework.

The aim of this analysis was to test the hypothesis that imputed rare and low-frequency variants in previously identified candidate genes are associated with celiac disease. The results from this study provide evidence that imputed rare and low-frequency variants from nearly fifty genes (mostly within the MHC region) are associated with celiac disease in 26 genes, four of which have previously been identified in previous studies. Further investigation with directly genotyped or resequenced independent samples will be required to validate the novel associations.

Chapter 7: Discussion and Conclusions

This dissertation project was undertaken to further the understanding of the genetic determinants of celiac disease. More specifically, this project has examined the existing literature on the genetics of celiac disease and aimed to investigate the validity of loci identified in previous celiac disease GWAS and interrogate the highly variable MHC region for evidence of associated loci that have been masked by the causal *HLA* alleles. This project has also investigated the use of NGS data to discover association signals amongst rare and low-frequency variants. Finally, this project aimed to investigate imputation of rare and low-frequency variants into a large-scale GWAS dataset.

Chapter 3

This study evaluated the relative performance of several meta-analysis methods to determine optimal methods for combining results from GWASs in the presence of between-study heterogeneity. Past studies have studied the role of heterogeneity in GWAS meta-analysis[94,237] and the development of novel meta-analysis models to better adjust for the observed heterogeneity[100,101]. While previous large-scale GWASs of celiac disease[72,73] have performed de facto meta-analyses of several diverse collections of samples by pooling data together that is equivalent to a fixed-effects meta-analysis, a critical investigation of between-study heterogeneity had not been performed.

In this study, the between-study heterogeneity at each variant between the sample collections in each previous celiac disease GWAS[72,73] was estimated and then used in the implementation of the novel random-effects based models that have recently been developed by Han et al.[100,101]. When the between-study heterogeneity is properly accounted for, one of the loci that had been significantly associated in the Dubois et al.[72] study, *RUNX3*, is no longer significant under the novel random-effects models when only considering the p-value of the association. However, a new plotting framework by Han et al.[101] that utilizes their new test statistic, the m-value, which represents the posterior probability of effect existence in a given collection, provided evidence that the highly

heterogeneous variant in *RUNX3* has a high posterior probability of effect existence in several collections. In the re-analysis of the Trynka et al.[73] data, all of the variants that were significant in the original study were significant again when between-study heterogeneity was accounted for by the novel random-effects models. In the presence of moderate between-study heterogeneity, one of the novel random-effects models, the binary-effects model yielded marginally lower p-values than the pooling method employed in the original study at three variants.

While some of the collections were genotyped in different laboratories and may still be subject to some laboratory-level bias, a bias due to genotyping platform is not likely to exist because all samples were genotyped on Illumina genotyping platforms. This study also lacks independent samples to further validate any of the novel signals from the Trynka et al. study. However, this limitation is present in many meta-analyses of GWAS because the rationale for performing a meta-analysis is to avoid genotyping new samples in the first place. Finally, although the I^2 statistic is based on Cochran's Q statistic and truncates to zero if $Q < (k - 1)$, since k never goes above 12 in either re-analysis.

Chapter 4

This study identified novel loci associated with celiac disease within the classical MHC region. The associations at these novel loci were found to be independent of the *HLA-DQA1* and *HLA-DQB1* high-risk alleles. This study builds upon the North American replication GWAS performed by Garner et al.[71] and was the first study to fine-map the extended MHC region in celiac disease cases and controls by implementing a novel statistical approach to adjust for the high-risk *HLA* alleles and grouping of variants at a locus by linkage disequilibrium. It is noteworthy that none of the four variants found to be independently associated with celiac disease are within exonic regions or known to have some functional importance in the development of celiac disease. However, three of these SNPs have previously been implicated in GWAS of related autoimmune disorders such as type 1

diabetes mellitus, systemic lupus erythematosus, and psoriasis. Notably, the most significant SNP, rs9357152, was significant in the first celiac disease GWAS[68] and highly significant in a type 1 diabetes mellitus GWAS. These pleiotropic variants strengthen the argument that autoimmune disorders such as type 1 diabetes mellitus, celiac disease, Crohn's disease, systemic lupus erythematosus, and rheumatoid arthritis share a common genetic background[80,238–240].

One of the limitations of this study was the lack of an independent population outside of the North American sample to attempt a replication of the results. Also, as the high recombination rates and complex LD patterns within the MHC cannot be completely adjusted for by statistical methods, it may be that the *HLA-DQA1* and *HLA-DQB1* high-risk alleles are slightly correlated with the four independent variants. However, given that the approach taken to capture the effects of the high-risk *HLA* genotypes was based on a highly sensitive and highly specific *HLA* typing method and a parsimonious modeling method, this explanation is not likely. While rare and low-frequency variants were not included in this study because they were not included in the SNP panel used in this study, this study has provided evidence that a future dense genotyping or resequencing study may discover associations with rare or low-frequency variants at or near the implicated loci as was demonstrated in the resequencing study of chapter 4.

Chapter 5

This is the first study that has used targeted NGS resequencing to discover rare or low-frequency variants associated with celiac disease. Nearly all previous GWAS[68,69,71–73] of celiac disease have been performed using microarray technology that focuses nearly exclusively on common variants with the exception of a recent, large-scale targeted exome sequencing study of autoimmune disorders[75] that includes rare and low-frequency variants. However, this study differs from the aforementioned exome sequencing study by resequencing not just the exonic regions but entire genes or genomic regions that have at least one variant associated with celiac disease in previous GWAS (i.e. includes

coding and non-coding regions). The burden test has been shown to be sufficiently powered to detect gene-level associations when the effects of the collapsed variants are unidirectional (i.e. either deleterious or protective). In this study, the rare and low-frequency burden tests revealed suggestive evidence of association at genes within the 12 non-MHC regions that are either close to genes implicated by common variants in previous GWAS (*RSPH3* and *PAPOLG*) or were previously found to have a variant significantly associated with celiac disease (*TAGAP*). In contrast to the burden test, the C-alpha test was specifically designed to test for a gene-level association in the presence of a mixture of deleterious, neutral, and protective variants.

While the limited sample size of around 500 samples limits the power to detect rare and low-frequency variants using single-variant tests, this study provided evidence that it is sufficiently powered to detect gene-level associations. It is also possible that the burden and C-alpha tests did not detect all of the gene-level associations of the rare and low-frequency variants and there may be a optimal method for detecting low-freq and rare variants in the regions that were re-sequenced. However, as there are now at least a dozen methods to test for rare and low-frequency associations, it was not feasible to evaluate all of the methods in the scope of this dissertation.

In future resequencing studies of celiac disease, there are a couple of obvious routes to take: one is to increase the number of samples that are re-sequenced and the second is perform whole-genome sequencing. Both of these future directions are becoming feasible as sequencing is now approaching the much sought after $1000 mark for a whole genome (Illumina has recently made the bold claim that it has reached this goal). Perhaps a more viable alternative in the near-future is to perform genotyping of all samples on another custom dense genotyping array that has the rare and low-frequency variants or, as explored in chapter 6, to impute the rare and low-frequency variants into an existing large-scale celiac disease GWAS dataset.

Chapter 6

In this study imputed rare and low-frequency variants were found to be associated with celiac disease. These associations were found at the gene and single-variant levels. This is the first study to impute rare and low-frequency variants identified in previous GWASs of celiac disease into a large-scale celiac GWAS dataset of approximately 12,000 cases and 12,000 controls. The regions imputed include 12 non-MHC regions and the classical MHC region. These results are important because they provide further evidence that rare and low-frequency variants that were not genotyped or sequenced in previous studies are involved in the etiology of celiac disease and an impetus for future investigation by direct genotyping or resequencing of associated loci in another large-scale celiac disease dataset. The evidence to directly genotype or re-sequence samples is provided by the gene-level replication of associations that were discovered in chapter 4, particularly the gene-level associations that were found in the MHC region as a majority of the genes yielding evidence of association in chapter 4 were again found to yield suggestive or significant evidence of association in this study. Furthermore, over two dozen additional genes provided evidence of association with celiac disease under the C-alpha test. Interestingly, whereas the burden test yielded no significant or suggestive evidence of association for genes within the MHC region in chapter 4 with the resequenced data, the rare and low-frequency burden tests of the imputed variants provided evidence of six genes associated with celiac disease within the MHC region, one of which was identified when only the NS variants were included in the analysis, indicating that there may be rare or low-frequency variants within the MHC region that are unidirectional and were not detected in chapter 4. When only the imputed NS variants were included in the rare and low-frequency C-alpha tests, subsets of the genes found to have significant or suggestive evidence of association in the analysis of all imputed variants yielded evidence of association. No additional genes were discovered when only the NS variants were included in the C-alpha tests. These observations provide evidence

that rare and low-frequency non-coding variants may play a significant role in the etiology and severity of celiac disease.

Of the single-marker test results, the most interesting signals were the very rare ($p = 0.004$) SNPs associated with celiac disease that are harbored in the non-coding regions of *C6orf10* and *BTNL2* as well as the highly significant and high effect size intronic SNPs from *TNXB* (rs17201553 with MAF of 0.011 and OR of 4.40 with $p = 3.08 \times 10^{-42}$) and *C6orf10* (rs9268233 with MAF of 0.029 and OR of 3.21 with $p = 1.17 \times 10^{-38}$). While the power analysis indicates that even this study is underpowered to detect rare variants with effect sizes < 2 with GWS, there is an indication that this study is well-powered to detect rare variants with effect sizes ≥ 2. The low-frequency SNPs from *HCP5* and *BAT2* that provided GWS evidence of association in the single-marker test along with the significant signals for the genes *GPSM*3, *HCP5*, and *PSORS1C1* under the rare and low-frequency C-alpha test of the imputed variants seem to strengthen the findings for common SNPs in a fine-mapping study of celiac disease conducted previously. In the previous fine-mapping study, the common SNPs from *HCP5* and *BAT2* also were associated with lower risk of disease (OR of 0.57 and 0.60, respectively). While the C-alpha test implemented in chapter 4 failed to find significant evidence of association between rare variants and celiac disease except in *ABCF1*, both the single-marker test and the gene-based burden and C-alpha tests of the imputed rare variants have provided GWS evidence of association with disease.

Imputation performance may be increased in future studies by combining multiple large-scale reference panels that are not yet available. While there is a small marginal loss in accuracy from the pre-phasing of sample haplotypes, it is noted by the authors of the imputation software used in this study that the time saved will outweigh the marginal loss in accuracy because the amount of time necessary for the imputation software to phase haplotypes in real-time is several orders of magnitude higher. It should be noted that the imputation performance (in terms of yielding well-

imputed genotypes) for the genes/regions that were imputed were comparable to the performance seen by Mägi et al.. While this study was focused on discovering imputed low-frequency and rare variants that are associated with celiac disease, a future study may focus on gene-based, genome-wide (i.e. not restricted to certain genes or regions) meta-analysis as implemented in an unpublished method such as RAREMETAL that was developed for the GSCAN project ("GWAS & Sequencing Consortium of Alcohol and Nicotine use"; http://gscan.sph.umich.edu). The RAREMETAL package was written to implement gene-based tests (burden based, variable threshold, SKAT) in an meta-analysis framework.

Conclusions

The results from chapter 3 reveal significant evidence that loci independent of the causal *HLA* alleles exist within the classical MHC region of chromosome 6 by accounting for both the effects of the high-risk *HLA* genotypes and the local LD patterns at associated loci and that further investigation of the MHC region is warranted. In chapter 4, the burden test and the C-alpha test provided significant evidence of association at the gene-level when rare and low-frequency variants are aggregated together and strengthens findings from previous GWAS and identified novel loci to investigate further, particularly in the MHC region. Chapter 5 evaluated FE and RE models of meta-analysis and provides evidence that if genotype data for multiple collections is available, a pooled analysis is sufficient, even in the presence of between-study heterogeneity and that most of the previous GWAS results remain significant after accounting for between-study heterogeneity in RE models. Chapter 5 also reveals that if only summary statistics are available, an RE-based meta-analysis may be necessary to account for between-study heterogeneity. Finally, chapter 6 provided evidence that high-quality imputation of rare and low-frequency variants using the latest imputation algorithms and reference panel is not only feasible but was able to replicate nearly all of the gene-

level associations from chapter 4 and even detect more novel loci that are associated with celiac disease.

References

1. Lander ES. Initial sequencing and analysis of the human genome. Nature. 2001 February;409(6822):860–921.

2. Botstein D, Risch N. Discovering genotypes underlying human phenotypes: past successes for mendelian disease, future approaches for complex disease. Nature genetics. 2003 March;33 Suppl(march):228–37.

3. Kerem B, Rommens J, Buchanan J. Identification of the Cystic Fibrosis Gene: Genetic Analysis. Science. 1989;245(4922):1073–1080.

4. Macdonald ME, Ambrose CM, Duyao MP, Myers RH, Lin C, Srinidhi L, Barnes G, Taylor SA, James M, Groat N, et al. A novel gene containing a trinucleotide repeat that is expanded and unstable on Huntington's disease chromosomes. Cell. 1993;72:971–983.

5. Hardy J, Singleton A. Genomewide association studies and human disease. The New England journal of medicine. 2009 April 23;360(17):1759–68.

6. Risch NJ. in the new millennium. 2000;405(JUNE).

7. Altmüller J, Palmer LJ, Fischer G, Scherb H, Wjst M. Genomewide scans of complex human diseases: true linkage is hard to find. American journal of human genetics. 2001 November;69(5):936–50.

8. Collins FS, Guyer MS, Chakravarti A. Variations on a Theme: Cataloging Human DNA Sequence Variation. Science. 1997 November 28;278(5343):1580–1581.

9. Collins FS, Green ED, Guttmacher AE, Guyer MS. A vision for the future of genomics research. Nature. 2003 April 24;422(6934):835–47.

10. Klein RJ, Zeiss C, Chew EY, Tsai J-Y, Sackler RS, Haynes C, Henning AK, SanGiovanni JP, Mane SM, Mayne ST, et al. Complement Factor H Polymorphism in Age-Related Macular Degeneration. Science. 2005 April 15;308(5720):385–389.

11. Barrett JC, Clayton DG, Concannon P, Akolkar B, Cooper JD, Erlich H a, Julier C, Morahan G, Nerup J, Nierras C, et al. Genome-wide association study and meta-analysis find that over 40 loci affect risk of type 1 diabetes. Nature genetics. 2009 June;41(6):703–7.

12. Smerdel-Ramoya a, Finholt C, Lilleby V, Gilboe I-M, Harbo HF, Maslinski S, Førre Ø, Thorsby E, Lie B a. Systemic lupus erythematosus and the extended major histocompatibility complex--evidence for several predisposing loci. Rheumatology (Oxford, England). 2005 November;44(11):1368–73.

13. Zeggini E, Scott LJ, Saxena R, Voight BF, Marchini JL, Hu T, de Bakker PI, Abecasis GR, Almgren P, Andersen G, et al. Meta-analysis of genome-wide association data and large-scale

replication identifies additional susceptibility loci for type 2 diabetes. Nature Genetics. 2008;40(5):638–645.

14. Kathiresan S, Melander O, Guiducci C, Surti A, Burtt NP, Rieder MJ, Cooper GM, Roos C, Voight BF, Havulinna AS, et al. Six new loci associated with blood low-density lipoprotein cholesterol, high-density lipoprotein cholesterol or triglycerides in humans. Nature genetics. 2008 February;40(2):189–97.

15. Pritchard JK. Are rare variants responsible for susceptibility to complex diseases? American journal of human genetics. 2001 July;69(1):124–37.

16. Gibson G. Rare and common variants: twenty arguments. Nature Reviews Genetics. 2012 January 18;13(2):135–145.

17. Bodmer W, Bonilla C. Common and rare variants in multifactorial susceptibility to common diseases. Nature genetics. 2008 June;40(6):695–701.

18. Smyth DJ, Cooper JD, Bailey R, Field S, Burren O, Smink LJ, Guja C, Ionescu-Tirgoviste C, Widmer B, Dunger DB, et al. A genome-wide association study of nonsynonymous SNPs identifies a type 1 diabetes locus in the interferon-induced helicase (IFIH1) region. Nature genetics. 2006 June;38(6):617–9.

19. Ueda H, Howson JMM, Esposito L, Heward J, Snook H, Chamberlain G, Rainbow DB, Hunter KMD, Smith AN, Di Genova G, et al. Association of the T-cell regulatory gene CTLA4 with susceptibility to autoimmune disease. Nature. 2003 May 29;423(6939):506–11.

20. Dahlman I, Eaves I a, Kosoy R, Morrison VA, Heward J, Gough SCL, Allahabadia A, Franklyn J a, Tuomilehto J, Tuomilehto-Wolf E, et al. Parameters for reliable results in genetic association studies in common disease. Nature genetics. 2002 February;30(2):149–50.

21. Manly KF, Nettleton D, Hwang JTG. Genomics, prior probability, and statistical tests of multiple hypotheses. Genome research. 2004 June;14(6):997–1001.

22. Wang WYS, Barratt BJ, Clayton DG, Todd J a. Genome-wide association studies: theoretical and practical concerns. Nature reviews. Genetics. 2005 February;6(2):109–18.

23. Kruglyak L. Prospects for whole-genome linkage disequilibrium mapping of common disease genes. Nature genetics. 1999 June;22(2):139–44.

24. Cortes A, Field J, Glazov E a, Hadler J, Stankovich J, Brown M a. Resequencing and fine-mapping of the chromosome 12q13-14 locus associated with multiple sclerosis refines the number of implicated genes. Human molecular genetics. 2013 June 1;22(11):2283–92.

25. Seddon JM, Yu Y, Miller EC, Reynolds R, Tan PL, Gowrisankar S, Goldstein JI, Triebwasser M, Anderson HE, Zerbib J, et al. Rare variants in CFI , C3 and C9 are associated with high risk of advanced age-related macular degeneration. Nature Genetics. 2013;45(11).

26. Zhan X, Larson DE, Wang C, Koboldt DC, Sergeev Y V, Fulton RS, Fulton LL, Fronick CC, Branham KE, Bragg-gresham J, et al. Identification of a rare coding variant in complement 3 associated with age-related macular degeneration. Nature Publishing Group. 2013;45(11):1375–1379.

27. Cardinale C, Wei Z, Panossian S. Targeted resequencing identifies defective variants of Decoy Receptor 3 in pediatric-onset inflammatory bowel disease. Genes and immunity. 2013;14(7):447–452.

28. Cohen JC, Kiss RS, Pertsemlidis A, Marcel YL, McPherson R, Hobbs HH. Multiple rare alleles contribute to low plasma levels of HDL cholesterol. Science. 2004 August 6;305(5685):869–72.

29. Moonesinghe R, Khoury MJ, Liu T, Ioannidis JP a. Required sample size and nonreplicability thresholds for heterogeneous genetic associations. Proceedings of the National Academy of Sciences of the United States of America. 2008 January 15;105(2):617–22.

30. Chapman K, Ferreira T, Morris A. Defining the power limits of genome-wide association scan meta-analyses. Genetic epidemiology. 2011;35(8):781–789.

31. Ioannidis JP a, Trikalinos T a, Khoury MJ. Implications of small effect sizes of individual genetic variants on the design and interpretation of genetic association studies of complex diseases. American journal of epidemiology. 2006 October 1;164(7):609–14.

32. Evangelou E, Valdes AM, Kerkhof HJM, Styrkarsdottir U, Zhu Y, Meulenbelt I, Lories RJ, Karassa FB, Tylzanowski P, Bos SD, et al. Meta-analysis of genome-wide association studies confirms a susceptibility locus for knee osteoarthritis on chromosome 7q22. Annals of the rheumatic diseases. 2011 February;70(2):349–55.

33. Todd J. Statistical false positive or true disease pathway? Nature genetics. 2006;38(7):5–7.

34. Panagiotou O a, Willer CJ, Hirschhorn JN, Ioannidis JP a. The Power of Meta-Analysis in Genome-Wide Association Studies. Annual review of genomics and human genetics. 2013 May 24;(May):1–25.

35. Zeggini E, Ioannidis JP a. Meta-analysis in genome-wide association studies. Pharmacogenomics. 2009 February;10(2):191–201.

36. Amundadottir LT, Sulem P, Gudmundsson J, Helgason A, Baker A, Agnarsson B a, Sigurdsson A, Benediktsdottir KR, Cazier J-B, Sainz J, et al. A common variant associated with prostate cancer in European and African populations. Nature genetics. 2006 June;38(6):652–8.

37. Vella A, Cooper JD, Lowe CE, Walker N, Nutland S, Widmer B, Jones R, Ring SM, McArdle W, Pembrey ME, et al. Localization of a type 1 diabetes locus in the IL2RA/CD25 region by use of tag single-nucleotide polymorphisms. American journal of human genetics. 2005 May;76(5):773–9.

38. Evangelou E, Ioannidis JP a. Meta-analysis methods for genome-wide association studies and beyond. Nature reviews. Genetics. 2013 June;14(6):379–89.

39. Pfeiffer RM, Gail MH, Pee D. On Combining Data From Genome-Wide Association Studies to Discover Disease-Associated SNPs. Statistical Science. 2009 November;24(4):547–560.

40. Cochran W. The Combination of Estimates from Different Experiments. Biometrics. 1954;10(1):101–129.

41. Jiao S, Hsu L, Hutter CM, Peters U. The use of imputed values in the meta-analysis of genome-wide association studies. Genetic epidemiology. 2011 November;35(7):597–605.

42. DerSimonian R, Laird N. Meta-analysis in clinical trials. Controlled clinical trials. 1986 September;7(3):177–88.

43. Pereira T V, Patsopoulos N a, Salanti G, Ioannidis JP a. Discovery properties of genome-wide association signals from cumulatively combined data sets. American journal of epidemiology. 2009 November 15;170(10):1197–206.

44. Losowsky MS. A history of coeliac disease. Digestive diseases (Basel, Switzerland). 2008 January;26(2):112–20.

45. Ferguson A, Arranz E, O'mahony S. Clinical and pathological spectrum of coeliac disease - active , silent , latent , potential. Gut. 1993;34:150–151.

46. Hopman EGD, le Cessie S, von Blomberg BME, Mearin ML. Nutritional management of the gluten-free diet in young people with celiac disease in The Netherlands. Journal of pediatric gastroenterology and nutrition. 2006 July;43(1):102–8.

47. Wahab PJ, Meijer JWR, Mulder CJJ. Histologic follow-up of people with celiac disease on a gluten-free diet: slow and incomplete recovery. American journal of clinical pathology. 2002 September;118(3):459–63.

48. Tack GJ, Verbeek WHM, Schreurs MWJ, Mulder CJJ. The spectrum of celiac disease: epidemiology, clinical aspects and treatment. Nature reviews. Gastroenterology & hepatology. 2010 April;7(4):204–13.

49. Cummins AG, Roberts-Thomson IC. Prevalence of celiac disease in the Asia-Pacific region. Journal of gastroenterology and hepatology. 2009 August;24(8):1347–51.

50. Lohi S, Mustalahti K, Kaukinen K, Laurila K, Collin P, Rissanen H, Lohi O, Bravi E, Gasparin M, Reunanen a, et al. Increasing prevalence of coeliac disease over time. Alimentary pharmacology & therapeutics. 2007 November 1;26(9):1217–25.

51. Dubé C, Rostom A, Sy R, Cranney A, Saloojee N, Garritty C, Sampson M, Zhang L, Yazdi F, Mamaladze V, et al. The prevalence of celiac disease in average-risk and at-risk Western European populations: A systematic review. Gastroenterology. 2005 April;128(4):S57–S67.

52. Catassi C, Rätsch IM, Fabiani E, Rossini M, Bordicchia F, Candela F, Coppa G V, Giorgi PL. Coeliac disease in the year 2000: exploring the iceberg. Lancet. 1994 January 22;343(8891):200–3.

53. Jacobson D, Gange S. Epidemiology and Estimated Population Burden of Selected Autoimmune Diseases in the United States. Clinical immunology and …. 1997;84(3):223–243.

54. Hoffenberg EJ, MacKenzie T, Barriga KJ, Eisenbarth GS, Bao F, Haas JE, Erlich H, Bugawan Tl TL, Sokol RJ, Taki I, et al. A prospective study of the incidence of childhood celiac disease. The Journal of pediatrics. 2003 September;143(3):308–14.

55. Catassi C, Bearzi I, Holmes GKT. Association of celiac disease and intestinal lymphomas and other cancers. Gastroenterology. 2005 April;128(4):S79–S86.

56. Rostami K, Kerckhaert J, Tiemessen R, von Blomberg BM, Meijer JW, Mulder CJ. Sensitivity of antiendomysium and antigliadin antibodies in untreated celiac disease: disappointing in clinical practice. The American journal of gastroenterology. 1999 April;94(4):888–94.

57. Rostom A, Dubé C, Cranney A, Saloojee N, Sy R, Garritty C, Sampson M, Zhang L, Yazdi F, Mamaladze V, et al. The diagnostic accuracy of serologic tests for celiac disease: A systematic review. Gastroenterology. 2005 April;128(4):S38–S46.

58. Sollid L, Thorsby E. HLA susceptibility genes in celiac disease: genetic mapping and role in pathogenesis. Gastroenterology. 1993 September;105(3):910–922.

59. Sollid LM. Coeliac disease: dissecting a complex inflammatory disorder. Nature Reviews Immunology. 2002;2(9):647–655.

60. Lewis NR, Scott BB. Meta-analysis: deamidated gliadin peptide antibody and tissue transglutaminase antibody compared as screening tests for coeliac disease. Alimentary pharmacology & therapeutics. 2010 January;31(1):73–81.

61. Walker-Smith J, Guandalini S, Report S. Revised criteria for diagnosis of coeliac disease. Archives of disease in …. 1990:909–911.

62. Mohamed BM, Feighery C, Coates C, O'Shea U, Delaney D, O'Briain S, Kelly J, Abuzakouk M. The absence of a mucosal lesion on standard histological examination does not exclude diagnosis of celiac disease. Digestive diseases and sciences. 2008 January;53(1):52–61.

63. Plot L, Amital H. Infectious associations of Celiac disease. Autoimmunity reviews. 2009 February;8(4):316–9.

64. Stene LC, Honeyman MC, Hoffenberg EJ, Haas JE, Sokol RJ, Emery L, Taki I, Norris JM, Erlich H a, Eisenbarth GS, et al. Rotavirus infection frequency and risk of celiac disease autoimmunity in early childhood: a longitudinal study. The American journal of gastroenterology. 2006 October;101(10):2333–40.

65. Belzen M Van, Meijer J, Sandkuijl L, Houwen RHJ, Wijmenga C. A major non-HLA locus in celiac disease maps to chromosome 19. Gastroenterology. 2003;5085(03):1032–1041.

66. Ding YC, Weizman Z, Yerushalmi B, Elbedour K, Garner CP, Neuhausen SL. An autosomal genome-wide screen for celiac disease in Bedouin families. Genes and immunity. 2008 January;9(1):81–6.

67. Garner CP, Ding YC, Steele L, Book L, Leiferman K, Zone JJ, Neuhausen SL. Genome-wide linkage analysis of 160 North American families with celiac disease. Genes and immunity. 2007 March;8(2):108–14.

68. Van Heel DA, Franke L, Hunt KA, Gwilliam R, Zhernakova A, Inouye M, Wapenaar MC, Barnardo MCNM, Bethel G, Holmes GKT, et al. A genome-wide association study for celiac disease identifies risk variants in the region harboring IL2 and IL21. Nature Genetics. 2007 July;39(7):827–829.

69. Hunt KA, Zhernakova A, Turner G, Heap GAR, Franke L, Bruinenberg M, Romanos J, Dinesen LC, Ryan AW, Panesar D, et al. Newly identified genetic risk variants for celiac disease related to the immune response. Nature Genetics. 2008;40(4):395–402.

70. Trynka G, Zhernakova A, Romanos J, Franke L, Hunt KA, Turner G, Bruinenberg M, Heap GA, Platteel M, Ryan AW, et al. Coeliac disease-associated risk variants in TNFAIP3 and REL implicate altered NF- B signalling. Gut. 2009 August;58(8):1078–1083.

71. Garner CPP, Murray JAA, Ding YCC, Tien Z, Van Heel DAA, Neuhausen SLL. Replication of celiac disease UK genome-wide association study results in a US population. Hum. Mol. Genet. 2009 November;18(21):4219.

72. Dubois PCA, Trynka G, Franke L, Hunt KA, Romanos J, Curtotti A, Zhernakova A, Heap GAR, Ádány R, Aromaa A, et al. Multiple common variants for celiac disease influencing immune gene expression. Nature Genetics. 2010;42(4):295–302.

73. Trynka G, Hunt K a, Bockett N a, Romanos J, Mistry V, Szperl A, Bakker SF, Bardella MT, Bhaw-Rosun L, Castillejo G, et al. Dense genotyping identifies and localizes multiple common and rare variant association signals in celiac disease. Nature Genetics. 2011 November 6;43(12):1193–1201.

74. Ahn R, Ding YC, Murray J, Fasano A, Green PHR, Neuhausen SL, Garner C. Association Analysis of the Extended MHC Region in Celiac Disease Implicates Multiple Independent Susceptibility Loci. PLoS ONE. 2012 January;7(5):e36926.

75. Hunt K a., Mistry V, Bockett N a., Ahmad T, Ban M, Barker JN, Barrett JC, Blackburn H, Brand O, Burren O, et al. Negligible impact of rare autoimmune-locus coding-region variants on missing heritability. Nature. 2013 May 22;498(7453):232–235.

76. Trynka G, Hunt K a, Bockett N a, Romanos J, Mistry V, Szperl A, Bakker SF, Bardella MT, Bhaw-Rosun L, Castillejo G, et al. Dense genotyping identifies and localizes multiple common and rare variant association signals in celiac disease. Nature Genetics. 2011 November;43(12):1193–1201.

77. Koskinen LLE, Einarsdottir E, Dukes E, Heap G a R, Dubois P, Korponay-Szabo IR, Kaukinen K, Kurppa K, Ziberna F, Vatta S, et al. Association study of the IL18RAP locus in three European populations with coeliac disease. Human molecular genetics. 2009 March 15;18(6):1148–55.

78. Romanos J, Barisani D, Trynka G, Zhernakova a, Bardella MT, Wijmenga C. Six new coeliac disease loci replicated in an Italian population confirm association with coeliac disease. Journal of medical genetics. 2009 January;46(1):60–3.

79. Megiorni F, Mora B, Bonamico M, Barbato M, Montuori M, Viola F, Trabace S, Mazzilli MC. HLA-DQ and susceptibility to celiac disease: evidence for gender differences and parent-of-origin effects. The American journal of gastroenterology. 2008 April;103(4):997–1003.

80. Coenen MJH, Trynka G, Heskamp S, Franke B, van Diemen CC, Smolonska J, van Leeuwen M, Brouwer E, Boezen MH, Postma DS, et al. Common and different genetic background for rheumatoid arthritis and coeliac disease. Human molecular genetics. 2009 November 1;18(21):4195–203.

81. Koskinen LLE, Einarsdottir E, Korponay-Szabo IR, Kurppa K, Kaukinen K, Sistonen P, Pocsai Z, Széles G, Adány R, Mäki M, et al. Fine mapping of the CELIAC2 locus on chromosome 5q31-q33 in the Finnish and Hungarian populations. Tissue antigens. 2009 November;74(5):408–16.

82. Dema B, Martínez a, Fernández-Arquero M, Maluenda C, Polanco I, de la Concha EG, Urcelay E, Núñez C. Association of IL18RAP and CCR3 with coeliac disease in the Spanish population. Journal of medical genetics. 2009 September;46(9):617–9.

83. Walker-Smith J, Guandalini S. Revised criteria for diagnosis of coeliac disease. Archives of disease in …. 1990:909–911.

84. Tian C, Gregersen PK, Seldin MF. Accounting for ancestry: population substructure and genome-wide association studies. Human Molecular Genetics. 2008;17(R2):R143–R150.

85. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M, et al. The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. Genome Research. 2010;20(9):1297–1303.

86. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics (Oxford, England). 2009 July 15;25(14):1754–60.

87. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R. The Sequence Alignment/Map format and SAMtools. Bioinformatics (Oxford, England). 2009 August 15;25(16):2078–9.

88. Saxena R, Voight BF, Lyssenko V, Burtt NP, de Bakker PIW, Chen H, Roix JJ, Kathiresan S, Hirschhorn JN, Daly MJ, et al. Genome-wide association analysis identifies loci for type 2 diabetes and triglyceride levels. Science. 2007 June 1;316(5829):1331–6.

89. Barroso I, Luan J, Wheeler E, Whittaker P, Wasson J, Zeggini E, Weedon MN, Hunt S, Venkatesh R, Frayling TM, et al. Population-specific risk of type 2 diabetes conferred by HNF4A P2 promoter variants: a lesson for replication studies. Diabetes. 2008 November;57(11):3161–5.

90. Ellinor PT, Lunetta KL, Albert CM, Glazer NL, Ritchie MD, Smith A V, Arking DE, Müller-Nurasyid M, Krijthe BP, Lubitz S a, et al. Meta-analysis identifies six new susceptibility loci for atrial fibrillation. Nature Genetics. 2012 April 29;44(6):670–675.

91. Franke A, McGovern DPB, Barrett JC, Wang K, Radford-Smith GL, Ahmad T, Lees CW, Balschun T, Lee J, Roberts R, et al. Genome-wide meta-analysis increases to 71 the number of confirmed Crohn's disease susceptibility loci. Nat Genet. 2010 December;42(12):1118–1125.

92. McMahon FJ, Akula N, Schulze TG, Muglia P, Tozzi F, Detera-Wadleigh SD, Steele CJM, Breuer R, Strohmaier J, Wendland JR, et al. Meta-analysis of genome-wide association data identifies a risk locus for major mood disorders on 3p21.1. Nature genetics. 2010 February;42(2):128–31.

93. Evangelou E, Maraganore DM, Ioannidis JP a. Meta-analysis in genome-wide association datasets: strategies and application in Parkinson disease. PloS one. 2007 January;2(2):e196.

94. Ioannidis JP a, Patsopoulos N a, Evangelou E. Heterogeneity in meta-analyses of genome-wide association investigations. PloS one. 2007 January;2(9):e841.

95. Sutton AJ, Higgins JPT. Recent developments in meta-analysis. Statistics in medicine. 2008;27(June):625–650.

96. Thompson JR, Attia J, Minelli C. The meta-analysis of genome-wide association studies. Briefings in bioinformatics. 2011 May;12(3):259–69.

97. Mantel N, Haenszel W. Statistical Aspects of the Analysis of Data From Retrospective Studies of Disease. Journal of the National Cancer Institute. 1959;22(4).

98. Fay MP, Graubard BI, Freedman LS, Midthune DN. Conditional Logistic Regression with Sandwich Estimators : Application to a Meta-Analysis Conditional Logistic Regression with Sandwich Estimators : Application to a Meta-Analysis. Biometrics. 1998;54(1):195–208.

99. Thompson SG, Sharp SJ. Explaining heterogeneity in meta-analysis: a comparison of methods. Statistics in medicine. 1999 October 30;18(20):2693–708.

100. Han B, Eskin E. Random-effects model aimed at discovering associations in meta-analysis of genome-wide association studies. Kerr K, editor. American journal of human genetics. 2011 May 13;88(3):e1002555.

101. Han B, Eskin E. Interpreting Meta-Analyses of Genome-Wide Association Studies Kerr K, editor. PLoS Genetics. 2012 March 1;8(3):e1002555.

102. Higgins JPT, Thompson SG. Quantifying heterogeneity in a meta-analysis. Statistics in medicine. 2002 June 15;21(11):1539–58.

103. Huedo-Medina TB, Sánchez-Meca J, Marín-Martínez F, Botella J. Assessing heterogeneity in meta-analysis: Q statistic or I2 index? Psychological methods. 2006 June;11(2):193–206.

104. Begum F, Ghosh D, Tseng GC, Feingold E. Comprehensive literature review and statistical considerations for GWAS meta-analysis. Nucleic acids research. 2012 May;40(9):3777–84.

105. Stephens M, Balding DJ. Bayesian statistical methods for genetic association studies. Nature reviews. Genetics. 2009 October;10(10):681–90.

106. Tang H. Confronting ethnicity-specific disease risk. Nature genetics. 2006 January;38(1):13–5.

107. Xiao R, Boehnke M. Quantifying and correcting for the winner's curse in genetic association studies. Genetic epidemiology. 2009 July;33(5):453–62.

108. Kavvoura FK, Ioannidis JP a. Methods for meta-analysis in genetic association studies: a review of their potential and pitfalls. Human genetics. 2008 February;123(1):1–14.

109. Wang X, Chua H-X, Chen P, Ong RT-H, Sim X, Zhang W, Takeuchi F, Liu X, Khor C-C, Tay W-T, et al. Comparing methods for performing trans-ethnic meta-analysis of genome-wide association studies. Human molecular genetics. 2013 June 1;22(11):2303–11.

110. Fasano A, Berti I, Gerarduzzi T. Prevalence of celiac disease in at-risk and not-at-risk groups in the United States: a large multicenter study. Archives of internal …. 2003 February;163(3):286–292.

111. Murray JA, Dyke C Van, Plevak MF, Dierkhising RA, Zinsmeister AR, Melton III LJ. Trends in the identification and clinical features of celiac disease in a North American community, 1950–2001. Clinical Gastroenterology and Hepatology. 2003 January;1(1):19–27.

112. Rubio-Tapia A, Kyle R a, Kaplan EL, Johnson DR, Page W, Erdtmann F, Brantner TL, Kim WR, Phelps TK, Lahr BD, et al. Increased prevalence and mortality in undiagnosed celiac disease. Gastroenterology. 2009 July;137(1):88–93.

113. Collin P, Pukkala E, Reunala T. Malignancy and survival in dermatitis herpetiformis: a comparison with coeliac disease. Gut. 1996 April 1;38(4):528–530.

114. Cooper BT, Holmes GK, Cooke WT. Coeliac disease and immunological disorders. British medical journal. 1978 March 4;1(6112):537–9.

115. Kaukinen K, Collin P, Mykkanen A, Partanen J, Maki M, Salmi J. Celiac disease and autoimmune endocrinologic disorders. Digestive diseases and sciences. 1999;44(7):1428–1433.

116. Ventura A, Magazzù G, Greco L. Duration of exposure to gluten and risk for autoimmune disorders in patients with celiac disease. SIGEP Study Group for Autoimmune Disorders in Celiac Disease. Gastroenterology. 1999;117:297–303.

117. Falchuk Z, Strober W. HL-A Antigens and adult coeliac disease. Lancet. 1972;2:1310.

118. Stokes P, Holmes G, Asquith P. Histocompatibility antigens associated with adult coeliac disease. The Lancet. 1972:162–164.

119. Tosi R, Vismara D, Tanigaki N, Ferrara GB, Cicimarra F, Buffolano W, Follo D, Auricchio S. Evidence that celiac disease is primarily associated with a DC locus allelic specificity. Clinical immunology and immunopathology. 1983 September;28(3):395–404.

120. Greco L, Corazza G, Babron MC, Clot F, Fulchignoni-Lataud MC, Percopo S, Zavattari P, Bouguerra F, Dib C, Tosi R, et al. Genome search in celiac disease. American journal of human genetics. 1998 March;62(3):669–75.

121. King a L, Fraser JS, Moodie SJ, Curtis D, Dearlove a M, Ellis HJ, Rosen-Bronson S, Ciclitira PJ. Coeliac disease: follow-up linkage study provides further support for existence of a susceptibility locus on chromosome 11p11. Annals of human genetics. 2001 July;65(Pt 4):377–86.

122. Liu J, Juo S-H, Holopainen P, Terwilliger J, Tong X, Grunn A, Brito M, Green P, Mustalahti K, Mäki M, et al. Genomewide linkage analysis of celiac disease in Finnish families. American journal of human genetics. 2002 January;70(1):51–9.

123. Popat S, Bevan S, Braegger CP, Busch a, O'Donoghue D, Falth-Magnusson K, Godkin a, Hogberg L, Holmes G, Hosie KB, et al. Genome screening of coeliac disease. Journal of medical genetics. 2002 May 1;39(5):328–31.

124. Rioux JD, Karinen H, Kocher K, McMahon SG, Kärkkäinen P, Janatuinen E, Heikkinen M, Julkunen R, Pihlajamäki J, Naukkarinen A, et al. Genomewide search and association studies in a Finnish celiac disease population: Identification of a novel locus and replication of the HLA and CTLA4 loci. American journal of medical genetics. Part A. 2004 November 1;130A(4):345–50.

125. Belzen M Van, Meijer J, Sandkuijl L. A major non-HLA locus in celiac disease maps to chromosome 19. Gastroenterology. 2003;5085(03):1032–1041.

126. Zhong F, McCombs C, Olson J. An autosomal screen for genes that predispose to celiac disease in the western counties of Ireland. Nature …. 1996;14:329–333.

127. Sollid L, Lundin K, Lundin H, Sjostrom H, Molberg O. HLA-DQ molecules, peptides and T cells in coeliac disease. In: Proceedings of the International Symposium on Coeliac Disease. ; 1997. pp. 265–274.

128. Hall RP, Sanders ME, Duquesnoy RJ, Katz SI, Shaw S. Alterations in HLA-DP and HLA-DQ Antigen Frequency in Patients with Dermatitis Herpetiformis. Journal of Investigative Dermatology. 1989;93(4):501–505.

129. Park MS, Terasaki PI, Ahmed AR, Zone J. The 90% incidence of HLA antigen (Te24) in dermatitis herpetiformis. Tissue Antigens. 1983;22(4):263–266.

130. Sachs JA, Awad J, McCloskey D, Navarrete C, Festenstein H, Elliot E, Walker-Smith JA, Griffiths CE, Leonard JN, Fry L. Different HLA associated gene combinations contribute to susceptibility for coeliac disease and dermatitis herpetiformis. Gut. 1986;27(5):515–520.

131. Balas A, Vicario JL, Zambrano A, Acuña D, García-Novo D. Absolute linkage of celiac disease and dermatitis herpetiformis to HLA-DQ. Tissue Antigens. 1997 July;50(1):52–56.

132. Bouguerra F, Babron MC, Eliaou JF, Debbabi A, Clot J, Khaldi F, Greco L, Clerget-Darpoux F. Synergistic effect of two HLA heterodimers in the susceptibility to celiac disease in Tunisia. Genetic Epidemiology. 1997;14(4):413–422.

133. Polvi A, Arranz E, Fernandez-Arquero M, Collin P, Mäki M, Sanz A, Calvo C, Maluenda C, Westman P, de la Concha EG, et al. HLA-DQ2-Negative Celiac Disease in Finland and Spain. Human Immunology. 1998 March;59(3):169–175.

134. Doherty DG, Vaughan RW, Donaldson PT, Mowat a P. HLA DQA, DQB, and DRB genotyping by oligonucleotide analysis: distribution of alleles and haplotypes in British caucasoids. Human immunology. 1992 May;34(1):53–63.

135. Lango A, Lindblom B. HLA DQA-DQB HAPLOTYPES IN A SWEDISH POPULATION. International Journal of Immunogenetics. 1993;20(6):453–460.

136. Mazzilli MC, Ferrante P, Mariani P, Martone E, Petronzelli F, Triglione P, Bonamico M. A study of Italian pediatric celiac disease patients confirms that the primary HLA association is to the DQ(alpha 1*0501, beta 1*0201) heterodimer. Human immunology. 1992 March;33(2):133–9.

137. Barcellos LF, May SL, Ramsay PP, Quach HL, Lane J a, Nititham J, Noble J a, Taylor KE, Quach DL, Chung S a, et al. High-density SNP screening of the major histocompatibility complex in systemic lupus erythematosus demonstrates strong evidence for independent susceptibility regions. PLoS genetics. 2009 October;5(10):e1000696.

138. Brown WM, Pierce J, Hilner JE, Perdue LH, Lohman K, Li L, Venkatesh RB, Hunt S, Mychaleckyj JC, Deloukas P. Overview of the MHC fine mapping data. Diabetes, Obesity and Metabolism. 2009;11(s1):2–7.

139. Howson JMM, Walker NM, Clayton D, Todd JA, Diabetes Genetics Consortium. Confirmation of HLA class II independent type 1 diabetes associations in the major histocompatibility complex including HLA-B and HLA-A. Diabetes, Obesity and Metabolism. 2009;11:31–45.

140. Nejentsev S, Howson JMM, Walker NM, Szeszko J, Field SF, Stevens HE, Reynolds P, Hardy M, King E, Masters J, et al. Localization of type 1 diabetes susceptibility to the MHC class I genes HLA-B and HLA-A. Nature. 2007 December;450(7171):887–892.

141. Breiman L, Friedman J, Olshen R, Stone C. Classification and Regression Trees. Boca Raton: Chapman and Hall/CRC; 1984.

142. Hothorn T, Hornik K, Zeileis A. Unbiased Recursive Partitioning: A Conditional Inference Framework. Journal of Computational and Graphical Statistics. 2006 September 1;15(3):651–674.

143. Purcell S, Neale B, Toddbrown K, Thomas L, Ferreira M, Bender D, Maller J, Sklar P, Debakker P, Daly M. PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. The American Journal of Human Genetics. 2007;81(3):559–575.

144. Aulchenko YS, Ripke S, Isaacs A, van Duijn CM. GenABEL: an R library for genome-wide association analysis. Bioinformatics. 2007 May;23(10):1294–1296.

145. Katz KD, Rashtak S, Lahr BD, Melton LJ, Krause PK, Maggi K, Talley NJ, Murray JA. Screening for Celiac Disease in a North American Population: Sequential Serology and Gastrointestinal Symptoms. The American Journal of Gastroenterology. 2011 July;106(7):1333–1339.

146. Walker MM, Murray J a, Ronkainen J, Aro P, Storskrubb T, D'Amato M, Lahr B, Talley NJ, Agreus L. Detection of celiac disease and lymphocytic enteropathy by parallel serology and histopathology in a population-based study. Gastroenterology. 2010 July;139(1):112–9.

147. Feolo M, Fuller TC, Taylor M, Zone JJ, Neuhausen SL. A strategy for high throughput HLA-DQ typing. Journal of immunological methods. 2001 December 1;258(1-2):65–71.

148. Monsuur AJ, de Bakker PIW, Zhernakova A, Pinto D, Verduijn W, Romanos J, Auricchio R, Lopez A, van Heel D a, Crusius JB a, et al. Effective detection of human leukocyte antigen risk alleles in celiac disease using tag single nucleotide polymorphisms. PloS one. 2008 January;3(5):e2270.

149. Horton R, Wilming L, Rand V, Lovering RC, Bruford EA, Khodiyar VK, Lush MJ, Povey S, Talbot CC, Wright MW, et al. Gene map of the extended human MHC. Nature Reviews Genetics. 2004 December;5(12):889–899.

150. Burton PR, Clayton DG, Cardon LR, Craddock N, Deloukas P, Duncanson A, Kwiatkowski DP, McCarthy MI, Ouwehand WH, Samani NJ, et al. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls : Article : Nature. Nature. 2007 June 7;447(7145):661–678.

151. Julià A, Ballina J, Cañete JD, Balsa A, Tornero-Molina J, Naranjo A, Alperi-López M, Erra A, Pascual-Salcedo D, Barceló P, et al. Genome-wide association study of rheumatoid arthritis in the Spanish population: KLF12 as a risk locus for rheumatoid arthritis susceptibility. Arthritis & Rheumatism. 2008;58(8):2275–2286.

152. Reich DE, Lander ES. On the allelic spectrum of human disease. Trends in genetics : TIG. 2001 September;17(9):502–10.

153. Manolio T a, Collins FS. The HapMap and genome-wide association studies in diagnosis and therapy. Annual review of medicine. 2009 January;60:443–56.

154. Project G, Asia E, Africa S, Figs S, Tables S. An integrated map of genetic variation. 2012;135(V):0–9.

155. Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, Hunter DJ, McCarthy MI, Ramos EM, Cardon LR, Chakravarti A, et al. Finding the missing heritability of complex diseases. Nature. 2009;461(7265):747–753.

156. Cirulli ET, Goldstein DB. Uncovering the roles of rare variants in common disease through whole-genome sequencing. Nature reviews. Genetics. 2010 June;11(6):415–25.

157. Pritchard JK, Cox NJ. The allelic architecture of human disease genes: common disease - common variant ... or not? Human Molecular Genetics. 2002;11:2417–2423.

158. Bulmer MG. Maintenance of genetic variability by mutation - selection balance: a child's guide through the jungle. 1989.

159. Charlesworth B. Anecdotal , Historical and Critical Commentaries on Genetics. 2000;931(November):927–931.

160. Lynch M. Rate, molecular spectrum, and consequences of human mutation. Proceedings of the National Academy of Sciences of the United States of America. 2010 January 19;107(3):961–8.

161. Cargill M, Altshuler D, Ireland J, Sklar P, Ardlie K, Patil N, Shaw N, Lane CR, Lim EP, Kalyanaraman N, et al. Characterization of single-nucleotide polymorphisms in coding regions of human genes. Nature genetics. 1999 July;22(3):231–8.

162. Kryukov G V, Pennacchio L a, Sunyaev SR. Most rare missense alleles are deleterious in humans: implications for complex disease and association studies. American journal of human genetics. 2007 April;80(4):727–39.

163. Zhu Q, Ge D, Maia JM, Zhu M, Petrovski S, Dickson SP, Heinzen EL, Shianna K V, Goldstein DB. A genome-wide comparison of the functional properties of rare and common genetic variants in humans. American journal of human genetics. 2011 April 8;88(4):458–68.

164. Voelkerding K V, Dames S a, Durtschi JD. Next-generation sequencing: from basic research to diagnostics. Clinical chemistry. 2009 April;55(4):641–58.

165. Sanger F. DNA sequencing with chain-terminating inhibitors. Proceedings of the …. 1977;74(12):5463–5467.

166. Metzker ML. Emerging technologies in DNA sequencing. Genome research. 2005 December;15(12):1767–76.

167. Hert DG, Fredlake CP, Barron AE. Advantages and limitations of next-generation sequencing technologies: a comparison of electrophoresis and non-electrophoresis methods. Electrophoresis. 2008 December;29(23):4618–26.

168. Mardis ER. Anticipating the 1,000 dollar genome. Genome biology. 2006 January;7(7):112.

169. Schloss J. How to get genomes at one ten-thousandth the cost. Nature biotechnology. 2008;26(10):1113–1115.

170. Linnarsson S. Recent advances in DNA sequencing methods - general principles of sample preparation. Experimental cell research. 2010 May 1;316(8):1339–43.

171. Metzker ML. Sequencing technologies—the next generation. Nature Reviews Genetics. 2009;11(1):31–46.

172. Mardis ER. A decade's perspective on DNA sequencing technology. Nature. 2011 February 10;470(7333):198–203.

173. Nagarajan N, Pop M. Sequencing and genome assembly using next-generation technologies. Computational Biology. 2010;673:1–17.

174. Pop M, Salzberg S. Bioinformatics challenges of new sequencing technology. Trends in Genetics. 2008;24(3):142–149.

175. McPherson J. Next-generation gap. Nature Methods. 2009;6(11):2–5.

176. Flicek P, Birney E. Sense from sequence reads : methods for alignment and assembly. Nature methods. 2009;6(11).

177. Burrows M, Wheeler DJ. A block-sorting lossless data compression algorithm. Citeseer; 1994.

178. Langmead B, Trapnell C, Pop M, Salzberg SL. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome biology. 2009 January;10(3):R25.

179. Li R, Yu C, Li Y, Lam T-W, Yiu S-M, Kristiansen K, Wang J. SOAP2: an improved ultrafast tool for short read alignment. Bioinformatics (Oxford, England). 2009 August 1;25(15):1966–7.

180. Li H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv preprint arXiv:1303.3997. 2013:1–3.

181. Nielsen R, Paul JS, Albrechtsen A, Song YS. Genotype and SNP calling from next-generation sequencing data. Nature Reviews Genetics. 2011 June;12(6):443–51.

182. Ewing B, Green P. Base-calling of automated sequencer traces using phred. II. Error probabilities. Genome research. 1998 March;8(3):186–94.

183. Brockman W, Alvarez P, Young S, Garber M, Giannoukos G, Lee WL, Russ C, Lander ES, Nusbaum C, Jaffe DB. Quality scores and SNP detection in sequencing-by-synthesis systems. Genome research. 2008 May;18(5):763–70.

184. Goldstein DB, Allen A, Keebler J, Margulies EH, Petrou S, Petrovski S, Sunyaev S. Sequencing studies in human genetics: design and interpretation. Nature reviews. Genetics. 2013 July;14(7):460–70.

185. Asimit J, Zeggini E. Rare variant association analysis methods for complex traits. Annual review of genetics. 2010 January;44:293–308.

186. Li B, Leal SM. Methods for Detecting Associations with Rare Variants for Common Diseases : Application to Analysis of Sequence Data. Journal of Human Genetics. 2008:311–321.

187. Morris AP, Zeggini E. An evaluation of statistical approaches to rare variant analysis in genetic association studies. Genetic epidemiology. 2010 February;34(2):188–93.

188. Neale BM, Rivas M a, Voight BF, Altshuler D, Devlin B, Orho-Melander M, Kathiresan S, Purcell SM, Roeder K, Daly MJ. Testing for an unusual distribution of rare variants. PLoS genetics. 2011 March;7(3):e1001322.

189. Kiezun A, Garimella K, Do R, Stitziel NO, Neale BM, McLaren PJ, Gupta N, Sklar P, Sullivan PF, Moran JL, et al. Exome sequencing and the genetic basis of complex traits. Nature genetics. 2012 June;44(6):623–30.

190. Nelson MR, Wegmann D, Ehm MG, Kessner D, St Jean P, Verzilli C, Shen J, Tang Z, Bacanu S-A, Fraser D, et al. An abundance of rare functional variants in 202 drug target genes sequenced in 14,002 people. Science (New York, N.Y.). 2012 July 6;337(6090):100–4.

191. Fu W, O'Connor TD, Jun G, Kang HM, Abecasis G, Leal SM, Gabriel S, Altshuler D, Shendure J, Nickerson D a., et al. Analysis of 6,515 exomes reveals the recent origin of most human protein-coding variants. Nature. 2012 November 28:6–10.

192. Nejentsev S, Walker N, Riches D, Egholm M, Todd JA. Rare Variants of IFIH1, a Gene Implicated in Antiviral Responses, Protect Against Type 1 Diabetes. 2009;333(April):387–389.

193. Momozawa Y, Mni M, Nakamura K, Coppieters W, Almer S, Amininejad L, Cleynen I, Colombel J-F, de Rijk P, Dewit O, et al. Resequencing of positional candidates identifies low frequency IL23R coding variants protecting against inflammatory bowel disease. Nature genetics. 2011 January;43(1):43–7.

194. Ajay SS, Parker SCJ, Abaan HO, Fajardo KVF, Margulies EH. Accurate and comprehensive sequencing of personal genomes. Genome research. 2011 September;21(9):1498–505.

195. San Lucas FA, Wang G, Scheet P, Peng B. Integrated annotation and analysis of genetic variants from next-generation sequencing studies with variant tools. Bioinformatics (Oxford, England). 2012 February 1;28(3):421–2.

196. Sherry ST, Ward MH, Kholodov M, Baker J, Phan L, Smigielski EM, Sirotkin K. dbSNP: the NCBI database of genetic variation. Nucleic acids research. 2001 January 1;29(1):308–11.

197. Pruitt KD, Tatusova T, Brown GR, Maglott DR. NCBI Reference Sequences (RefSeq): current status, new features and genome annotation policy. Nucleic acids research. 2012 January;40(Database issue):D130–5.

198. Liu X, Jian X, Boerwinkle E. dbNSFP: a lightweight database of human nonsynonymous SNPs and their functional predictions. Human mutation. 2011 August;32(8):894–9.

199. Panoutsopoulou K, Tachmazidou I, Zeggini E. In search of low-frequency and rare variants affecting complex traits. Human molecular genetics. 2013 October 15;22(R1):R16–21.

200. Diogo D, Kurreeman F, Stahl E a, Liao KP, Gupta N, Greenberg JD, Rivas M a, Hickey B, Flannick J, Thomson B, et al. Rare, low-frequency, and common variants in the protein-coding sequence of biological candidate genes from GWASs contribute to risk of rheumatoid arthritis. American journal of human genetics. 2013 January 10;92(1):15–27.

201. Mägi R, Asimit JL, Day-Williams AG, Zeggini E, Morris AP. Genome-Wide Association Analysis of Imputed Rare Variants: Application to Seven Common Complex Diseases. Genetic epidemiology. 2012 September 5;796:785–796.

202. Ebersberger I, Metzler D, Schwarz C, Pääbo S. Genomewide comparison of DNA sequences between humans and chimpanzees. American journal of human genetics. 2002 June;70(6):1490–7.

203. Kenny EE, Pe'er I, Karban A, Ozelius L, Mitchell A a, Ng SM, Erazo M, Ostrer H, Abraham C, Abreu MT, et al. A genome-wide scan of Ashkenazi Jewish Crohn's disease suggests novel susceptibility loci. PLoS genetics. 2012 January;8(3):e1002559.

204. DePaolo RW, Abadie V, Tang F, Fehlner-Peach H, Hall J a, Wang W, Marietta E V, Kasarda DD, Waldmann T a, Murray J a, et al. Co-adjuvant effects of retinoic acid and IL-15 induce inflammatory immunity to dietary antigens. Nature. 2011 March 10;471(7337):220–4.

205. Hüe S, Mention J-J, Monteiro RC, Zhang S, Cellier C, Schmitz J, Verkarre V, Fodil N, Bahram S, Cerf-Bensussan N, et al. A direct role for NKG2D/MICA interaction in villous atrophy during celiac disease. Immunity. 2004 September;21(3):367–77.

206. Allegretti YL, Bondar C, Guzman L, Cueto Rua E, Chopita N, Fuertes M, Zwirner NW, Chirdo FG. Broad MICA/B expression in the small bowel mucosa: a link between cellular stress and celiac disease. PloS one. 2013 January;8(9):e73658.

207. Asimit JL, Zeggini E. Imputation of rare variants in next-generation association studies. Human heredity. 2012 January;74(3-4):196–204.

208. Pei Y-F, Zhang L, Li J, Deng H-W. Analyses and Comparison of Imputation-Based Association Methods. PLoS ONE. 2010 May;5(5):e10827.

209. Sung YJ, Wang L, Rankinen T, Bouchard C, Rao DC. Performance of genotype imputations using data from the 1000 Genomes Project. Human heredity. 2012 January;73(1):18–25.

210. Marchini J, Howie B. Genotype imputation for genome-wide association studies. Nature reviews. Genetics. 2010 July;11(7):499–511.

211. Howie B, Marchini J, Stephens M. Genotype imputation with thousands of genomes. G3 (Bethesda, Md.). 2011 November;1(6):457–70.

212. Sterne J, White I, Carlin J. Multiple imputation for missing data in epidemiological and clinical research : potential and pitfalls. BMJ: British Medical …. 2009;339(July):157–160.

213. Harel O, Zhou X. Multiple imputation-Review of theory, implementation and software. Statistics in medicine. 2006;(January):3057–3077.

214. Allison P. Missing Data. Sage Publications; 2002.

215. Rubin D. Multiple imputation after 18+ years. Journal of the American Statistical Association. 1996;91(434):473–489.

216. Marchini J, Howie B, Myers S, McVean G, Donnelly P. A new multipoint method for genome-wide association studies by imputation of genotypes. Nature genetics. 2007 July;39(7):906–13.

217. Stephens M, Donnelly P. Inference in molecular population genetics. Journal of the Royal Statistical Society: Series B (Statistical Methodology). 2000 November;62(4):605–635.

218. Fearnhead P, Donnelly P. Estimating recombination rates from population genetic data. Genetics. 2001 November;159(3):1299–318.

219. Rabiner LR. A tutorial on hidden Markov models and selected applications in speech recognition. Proceedings of the IEEE. 1989;77(2):257–286.

220. Howie BN, Donnelly P, Marchini J. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. PLoS Genet. 2009;5(6):e1000529.

221. Anon. A map of human genome variation from population-scale sequencing. Nature. 2010 October;467(7319):1061–1073.

222. Li Y, Willer CJ, Ding J, Scheet P, Abecasis GR. MaCH: using sequence and genotype data to estimate haplotypes and unobserved genotypes. Genetic epidemiology. 2010;34(8):816–834.

223. Browning BL, Browning SR. A unified approach to genotype imputation and haplotype-phase inference for large data sets of trios and unrelated individuals. American journal of human genetics. 2009 February;84(2):210–23.

224. Zheng H-F, Ladouceur M, Greenwood CMT, Richards JB. Effect of genome-wide genotyping and reference panels on rare variants imputation. Journal of genetics and genomics. 2012 October 20;39(10):545–50.

225. Myers S, Bottolo L, Freeman C, McVean G, Donnelly P. A fine-scale map of recombination rates and hotspots across the human genome. Science (New York, N.Y.). 2005 October 14;310(5746):321–4.

226. Delaneau O, Zagury J-F, Marchini J. Improved whole-chromosome phasing for disease and population genetic studies. Nature methods. 2013 January;10(1):5–6.

227. The International HapMap3 Consortium. Europe PMC Funders Group Integrating common and rare genetic variation in diverse human populations. 2011;467(7311):52–58.

228. Jostins L, Morley KI, Barrett JC. Imputation of low-frequency variants using the HapMap3 benefits from large, diverse reference sets. European journal of human genetics : EJHG. 2011 June;19(6):662–6.

229. Hinrichs a S, Karolchik D, Baertsch R, Barber GP, Bejerano G, Clawson H, Diekhans M, Furey TS, Harte R a, Hsu F, et al. The UCSC Genome Browser Database: update 2006. Nucleic acids research. 2006 January 1;34(Database issue):D590–8.

230. Howie B, Fuchsberger C, Stephens M, Marchini J, Abecasis GR. Fast and accurate genotype imputation in genome-wide association studies through pre-phasing. Nature Genetics. 2012 July 22;44(8):955–959.

231. Scott LJ, Mohlke KL, Bonnycastle LL, Willer CJ, Li Y, Duren WL, Erdos MR, Stringham HM, Chines PS, Jackson AU, et al. A genome-wide association study of type 2 diabetes in Finns detects multiple susceptibility variants. Science. 2007 June 1;316(5829):1341–5.

232. De Bakker PIWW, Ferreira MARR, Jia X, Neale BM, Raychaudhuri S, Voight BF. Practical aspects of imputation-driven meta-analysis of genome-wide association studies. Human Molecular Genetics. 2008 October 15;17(R2):R122–R128.

233. Skol A. Joint analysis is more efficient than replication-based analysis for two-stage genome-wide association studies - Nature Genetics. 2006 [cited 2010 June 15]. Available from: http://www.nature.com/ng/journal/v38/n2/abs/ng1706.html

234. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira M a R, Bender D, Maller J, Sklar P, de Bakker PIW, Daly MJ, et al. PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. The American Journal of Human Genetics. 2007 September;81(3):559–575.

235. Panagiotou O a, Ioannidis JP a. What should the genome-wide significance threshold be? Empirical replication of borderline genetic associations. International journal of epidemiology. 2012 February;41(1):273–86.

236. Albers PK, Abecasis GR, McCarthy MI, Gaulton KJ. Meta-imputation: a simple and flexible method to combine multiple reference panels for imputing genetic variants. In: American Society of Human Genetics. Boston, MA, USA: American Society of Human Genetics; 2013.

237. Ioannidis J, Patsopoulos N, Evangelou E. Uncertainty in heterogeneity estimates in meta-analyses. BMJ. 2007;335(November).

238. Zhernakova A, van Diemen CC, Wijmenga C. Detecting shared pathogenesis from the shared genetics of immune-related diseases. Nature reviews. Genetics. 2009 January;10(1):43–55.

239. Zhernakova A, Stahl E a, Trynka G, Raychaudhuri S, Festen E a, Franke L, Westra H-J, Fehrmann RSN, Kurreeman F a S, Thomson B, et al. Meta-analysis of genome-wide association studies in celiac disease and rheumatoid arthritis identifies fourteen non-HLA shared loci. PLoS genetics. 2011 March;7(2):e1002004.

240. Trynka G, Wijmenga C, van Heel DA. A genetic perspective on coeliac disease. Trends in Molecular Medicine. 2010 November;16(11):537–50.