EXPLORING THE GENETIC ARCHITECTURE OF COMPLEX DISEASES WITH GENOME-WIDE ASSOCIATION STUDIES

A Dissertation

Presented to the Faculty of the Graduate School

of Cornell University

In Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy

by

Diana Chang

August 2014

© 2014 Diana Chang

EXPLORING THE GENETIC ARCHITECTURE OF COMPLEX DISEASES WITH GENOME-WIDE ASSOCIATION STUDIES

Diana Chang, Ph. D.

Cornell University 2014

Over the past decade, the number of genome-wide association studies (GWAS) carried out has increased exponentially. These studies, mostly by investigating single nucleotide polymorphisms (SNPs), have discovered thousands of new loci associated to numerous complex diseases and traits, such as Crohn's Disease, Type-1 and Type-2 diabetes, height and body mass index. Unfortunately, there are several limitations to current GWAS. Firstly these newly discovered associations fail to explain all of the observed phenotypic variability attributed to genetic sources. This issue of missing heritability can be attributed to multiple sources such as rare variants, epigenetics and gene-gene interactions. Secondly, the majority of GWAS have not investigated the contribution of the sex chromosomes to complex disease. And thirdly, though comorbidity studies have well-established the overlap between some diseases, many initial GWAS focused on single phenotypes, and are only recently investigating the genetic overlap between various complex diseases (and traits). Here, we investigate and extend various aspects of GWAS to address these issues. First, we investigate the implication of rare or low frequency causal variants (SNPs with a minor allele frequency <5%) for GWAS and find that when diseases are caused by (unassayed) rare variants, the associated SNPs tend to lie further away than expected when diseases are caused by common variants. Second, we investigate the role of chromosome X in complex disease. The X chromosome was routinely ignored and mishandled in many GWAS, thus possibly explaining the lack of X-linked associations. Hence, we developed

an X-tailored pipeline and applied it to 16 datasets of autoimmune and immune-mediated disorders. We found several genes implicated in disease risk, some of which have sex-differentiated function. Finally, we developed a novel method, *disPCA*, that uses principal component analysis to investigate the shared genetics between various complex diseases and traits. Applying *disPCA* to 31 GWAS datasets, we found several pathways that may underlie shared pathogenesis between distinct diseases and traits. Though genotyping-based GWAS are being quickly replaced with sequencing-based association studies, the conclusions and tools developed here can also be applied to this new generation of data.

BIOGRAPHICAL SKETCH

Diana Chang was born and raised in New York City. She attended Barnard College from 2004 to 2008 and graduated *magna cum laude* with a Bachelor of Arts in Applied Mathematics and Philosophy. During her bachelors, she carried out research in the field of Paleoclimatology at the Lamont Doherty Earth Observatory. After college, Diana took a short break and ventured into corporate culture providing application support for traders at Legg Mason. She returned to her studies in 2009, enrolling in the Tri-Institutional Training Program of Computational Biology and Medicine, working in Dr. Alon Keinan's lab on a variety of issues in human and medical genetics.

Dedicated to my grandmother.

ACKNOWLEDGMENTS

I am grateful to and thankful of my advisor, Alon Keinan, for the wonderful opportunities he has given me and for his advice, patience and confidence. I am also grateful to past and present members of the Keinan Lab and for all the encouragement they have provided me. I would like to also thank Andrew Clark, Jason Mezey and Grégoire Altan-Bonnet for their support. Finally, I would like to thank my family, and my friends for being there throughout this journey.

~

The work presented here was supported by The Tri-Institutional Training Program in Computational Biology and Medicine, The Ellison Medical Foundation, NIH Grant U01HG005715, NIH Training Grant T32GM083937, and NIH Grant R01HG006849.

TABLE OF CONTENTS

FIGURES

TABLES

Chapter 1： Introduction

The rapid decline in cost of assaying genetic information ushered in the era of genome-wide association studies (GWAS). In its simplest form, GWAS consists of comparing genetic information at different loci between individuals with a disease (cases) and those without (controls). Usually assayed at single nucleotide polymorphisms (SNPs), associated SNPs represent a significant correlation between case or control status and the SNP itself. GWAS are also carried out on quantitative phenotypes, such as height, in a similar fashion. Associations do not necessarily point to causality though, as SNPs assayed by genotyping arrays serve as tag markers for other SNPs or different types of variants that they are correlated with, i.e. within the same linkage disequilibrium (LD) block. Thus other untyped variants within the same LD block as an associated SNP may be the true underlying genetic variant/s.

Over the past decade, over 1350 studies have been reported (Hindorff, MacArthur et al. 2013) with over 11,000 associated SNPs (Welter, MacArthur et al. 2014). Each association holds promise for narrowing the search field for disease mechanisms and options for treatment. For example, the association of complement factor H to age-related macular degeneration is now being investigated as a therapeutic target for the disease (Troutbeck, Al-Qureshi et al. 2012). In essence, GWAS serve as one possible initial step to understanding the biology of diseases that can further advance medicine and eventually lead to effective measures of healthcare (Green and Guyer 2011).

Despite this aforementioned promise and projected success, early GWAS were limited and were expanded to address the following issues: 1) the case of missing heritability 2) exploring the

contribution of chromosome X and 3) the focus on single phenotypes. Firstly it was surprising to find that discovered associations had yet to explain a significant portion of genetic variability for many traits and disease risks (the problem of missing heritability) (Manolio, Collins et al. 2009). It was postulated that there were genetic variants and types of associations that conventional GWAS data and methods missed. These included epigenetics, genetic interactions, structural variants and low frequency variants (variants with minor allele frequency –MAF<0.05). The case of rare or low frequency variants was particularly troublesome as one study suggested that if rare variants underlay signals of association (the case of "synthetic associations"), then the actual causal variants could be much further away from an association signal than expected given common causal variants (Dickson, Wang et al. 2010). This in turn could suggest that fine mapping studies (studies following up GWAS to narrow down the causal locus behind an association signal) needed to explore a larger genomic area to find the actual causal variant underlying an association. While the results presented in Dickson *et al.* suggested reevaluating fine mapping strategies, the study based their results on simulated data. In order to better understand the phenomenon of synthetic associations though, one would need to explore the phenomenon in human genetic data with actual patterns of LD between common and rare variants. Thus, using such data I refined our expectations regarding synthetic associations (Chapter 2).

Another potential source of missing heritability is the X chromosome. In humans, females have two copies of the X chromosome, while males have one. It has been suggested that chromosome X may play a role in sex-specific disorders and diseases such as many autoimmune disorders (Ober, Loisel et al. 2008; Libert, Dejager et al. 2010; Bianchi, Lleo et al. 2011; Quintero,

Amador-Patarroyo et al. 2012; Selmi, Brunetta et al. 2012), and it is also implicated in some Mendelian disorders (Hamosh, Scott et al. 2002; Hamosh, Scott et al. 2005). Surprisingly, little evidence for a role of the X chromosome in complex diseases exists from GWAS. As of 2013, less than 50 associations exist on the X chromosome (Hindorff, MacArthur et al. 2013). While this may reflect a biological phenomenon (e.g. few mutations on chromosome X are risk variants for complex diseases and traits), a review of the literature suggests otherwise. Namely, 67% of GWAS during 2011 alone neglected to analyze chromosome X (Wise, Gyi et al. 2013). This itself was likely due to differences in statistical tests needed for chromosome X than the autosomes (non-sex chromosomes). Given this apparent gap in the field of GWAS and vast amount of unanalyzed data, we have developed a statistical package to carry out X-wide association analysis (XWAS) and further applied it to a number of autoimmune disorders (Chapter 3).

In addition to cracking the case of missing heritability, the plain vanilla model of GWAS was also extended to explore the genetic overlap between phenotypes. As more GWAS were published, the overlapping associations between phenotypes became apparent, many of which supported known comorbidities and pleiotropies (Sirota, Schaub et al. 2009; Cotsapas, Voight et al. 2011; Sivakumaran, Agakov et al. 2011; Solovieff, Cotsapas et al. 2013). For example, type-1 diabetes and rheumatoid arthritis, two diseases with known comorbidity (Somers, Thomas et al. 2009), share 12 associations (Hindorff, MacArthur et al. 2013). Thus as increasing evidence came to light regarding the genetic overlap between phenotypes, many have extended GWAS to be carried out on more than one phenotype (Klei, Luca et al. 2008; Hartley, Monti et al. 2012; Andreassen, Thompson et al. 2013; Solovieff, Cotsapas et al. 2013; Andreassen, Harbo et al.

2014). Various methods exist –some aim to identify shared genetic loci, while others aim to identify the pairs or sets of diseases that share pathogenesis. Here, I have developed a principal component based method that elucidates sets of diseases sharing genetic pathogenesis and highlights pathways that may be enriched for genes underlying shared pathogenesis. I have further applied this method to a number of GWAS datasets spanning autoimmune, neurological, psychiatric and other disorders and traits (Chapter 4).

Chapter 2: Predicting signatures of "synthetic associations" and "natural associations" from empirical patterns of human genetic variation

(Chang and Keinan 2012)

2.1 Introduction

Recent years have seen a plethora of Genome-wide association studies (GWAS) finding thousands of common markers that are associated with hundreds of diseases and other traits. GWAS were initially founded on the Common Disease-Common Variant hypothesis (Reich and Lander 2001; Pritchard and Cox 2002; Iles 2008), which predicted that common complex diseases are most likely caused by a few common variants. As a consequence, the design of most GWAS consisted of genotyping common tag single nucleotide polymorphisms (SNPs) and comparing their allele frequencies between cases and controls. Some limitations of this design have been the topic of much recent discussion, with the gap between association and causality and the relatively small portion of heritable variation explained by associated markers drawing the most concern (Maher 2008; Frazer, Murray et al. 2009; Manolio, Collins et al. 2009; Eichler, Flint et al. 2010). Several hypotheses aiming to explain the missing heritability have been proposed, including the roles of structural variants, gene-gene interactions, gene-environment interactions, epigenetics, and complex inheritance (Maher 2008; Frazer, Murray et al. 2009; Manolio, Collins et al. 2009; Eichler, Flint et al. 2010). In addition, rare variants of relatively high penetrance contributing to disease risk (Pritchard 2001; Bodmer and Bonilla 2008) has also been suggested as a source of missing heritability since rare variants have not been directly observed in most GWAS, and they might be differently tagged by common markers (McCarthy, Abecasis et al. 2008; Cirulli and Goldstein 2010; Wang, Dickson et al. 2010).

Given this renewed interest in such variants, an investigation into their effect on GWAS association signals is warranted. A recent simulation-based study showed that rare causal variants can often create "synthetic associations," namely significant associations of common markers induced by the combined effect of one or more rare causal variants (Dickson, Wang et al. 2010). Dickson et al. further showed that a synthetically associated common marker could be substantially further away than expected had the underlying causal variant been common, and that synthetic associations are expected to be on average of lower minor allele frequency (MAF) than associations due to underlying common causal variants (Dickson, Wang et al. 2010). These predictions may partially explain why resequencing fine-mapping efforts, which are based on patterns of linkage disequilibrium (LD) of common variants, have often been unsuccessful in uncovering causal variants (McCarthy, Abecasis et al. 2008; McCarthy and Hirschhorn 2008; Dickson, Wang et al. 2010). While the development of new methods and study designs for associating rare causal variants is underway (Madsen and Browning 2009; Bansal, Libiger et al. 2010; Cirulli and Goldstein 2010; Han and Pan 2010; Hoffmann, Marini et al. 2010; Longmate, Larson et al. 2010; Oksenberg and Baranzini 2010; Price, Kryukov et al. 2010; Rosenberg, Huang et al. 2010; Takeuchi, Kobayashi et al. 2011; Wu, Lee et al. 2011), the predictions of Dickson et al. are influencing analyses of such studies, as well as the interpretation of traditional, genotyping-based GWAS (e.g. (Fellay, Thompson et al. 2010; Shatunov, Mok et al. 2010)).

A few instances of rare causal variants have already been well established (Cohen, Boerwinkle et al. 2006; Romeo, Pennacchio et al. 2007; Kathiresan, Willer et al. 2009), including the recently discovered, potentially rare causal variants in NOD2 that contribute to Crohn's disease risk

(Hugot, Chamaillard et al. 2001; Ogura, Bonen et al. 2001; Bonen and Cho 2003; The Wellcome Trust Case Control Consortium 2007). In this example, since an associated common marker in the same gene is in LD with at least two of the rare variants, it is possible that they contribute to the marker's association signal (The Wellcome Trust Case Control Consortium 2007), thus inducing a synthetic association. As only a few examples of rare causal variants contributing to complex disease are well established, the jury is still out on their prevalence and on how often they lead to synthetic associations, with several recent studies arguing that the phenomenon is not necessarily widespread (Orozco, Barrett et al. 2010; Anderson, Soranzo et al. 2011; Wray, Purcell et al. 2011). In light of this uncertainty, a detailed investigation of the signatures of synthetic associations and their implications is crucial for interpreting the results of genotyping-based GWAS and for considering the alternative of association studies based on whole-genome or whole-exome sequencing.

Two of the key questions with regards to "synthetic associations" are (1) what are the implications for the resequencing distance for fine-mapping of significant associations? and (2) how different is the MAF of synthetic associations from that of "natural associations" (i.e. associations where the underlying causal variants are common)? While these questions have been addressed in studies of simulated data (Dickson, Wang et al. 2010; Wray, Purcell et al. 2011), those simulations did not account for the nature of disease loci and risk variants, nor did they account for the specific nature of human genetic variation. In the former, it has been shown that the effect size and frequency of the disease variants can alter the power of the test (Chapman, Cooper et al. 2003). While, in the latter, the mark left by human evolutionary history on patterns of genetic variation can greatly influence the nature of significant association signals,

which we address in the present study. For example, when considering samples from European populations, which have been the populations of choice of most GWAS, it is crucial to account for their recent explosive population growth that has led to an inflation in the proportion of rare variants and to an altered haplotype and LD structure, as well as to account for the well-established effects of the earlier Out-of-Africa event on these genetic patterns (Tishkoff, Dietzsch et al. 1996; Dunning, Durocher et al. 2000; Reich, Cargill et al. 2001; Adams and Hudson 2004; Marth, Czabarka et al. 2004; Keinan, Mullikin et al. 2007; Keinan, Mullikin et al. 2009; Keinan and Clark 2012).

Here, we focus on the question of how empirical LD patterns can affect signals of "synthetic association" by investigating them in real human population genetic data. Through this, we aim to derive a better understanding of synthetic associations and their practical implications. Using empirical resequencing data, we randomly assume certain variants as increasing disease risk, determine cases and controls accordingly, and conduct an association study using genotyping data of the same individuals from arrays that have been employed in most GWAS. To illuminate and quantify signatures that are specific to "synthetic associations", we repeat the process for rare and common causal variants and contrast the characteristics of synthetic associations with those of natural associations.

We aim to elucidate how far associations are from the underlying causal variants, how their frequencies are distributed and, more importantly, how these different signatures alter the design of fine-mapping studies. To examine possible heterogeneity in these signatures across the genome and across populations with different evolutionary histories, we repeat the analysis for

several resequencing loci on different chromosomes and for two populations, one West African and one North European. The novelty of this study is in elucidating implications of synthetic associations and how they may affect fine-mapping strategies with the use of data that maintains LD patterns observed in human populations.

2.2 Results

To empirically investigate the signatures of "synthetic associations", we needed to examine scenarios in human genetic data where the presumed disease risk variants—rare or common—are known. Thus, we considered "disease loci" in the ENCODE regions that were sequenced as part of HapMap 3 (Altshuler, Gibbs et al. 2010). The advantage of using these resequencing data is that we could observe variants of much lower allele frequency that are also free of ascertainment biases, which plague genotyping arrays (Clark, Hubisz et al. 2005; Frazer, Ballinger et al. 2007; Keinan, Mullikin et al. 2007; Albrechtsen, Nielsen et al. 2010). Equipped with resequencing data for over 110 individuals in each population, we studied variants that were of frequency as low as 0.9% (after exclusion of singletons). We randomly assigned variants within each disease locus as being causal and considered individuals carrying any one of these variants to have elevated disease risk. We then probabilistically assigned individuals to be either cases or controls based on their assigned risk. To mimic the case of many rare variants of large effect size underlying synthetic associations, and to contrast it with that of a few common variants of moderately low effect sizes underlying natural associations, we investigated three scenarios: (i) 2 common causal variants with a genotypic relative risk (GRR) of 1.5, (ii) 5 and (iii) 9 rare causal variants with a genotypic relative risk of 3. We verified that our results are not an artifact of the number of

causal variants, as illustrated in the following, by comparing with a less realistic scenario of 5 common causal variants. We also considered a random assignment of cases and controls, which provides a null distribution in the absence of any risk alleles.

After obtaining a set of cases and a set of controls, we performed an association study using the genotyping array data for the same individuals from HapMap 3 (Altshuler, Gibbs et al. 2010), without considering any of the resequencing data in which disease loci have been emulated (Materials and Methods). This mimics the conditions and variant-type of actual genotyping-based GWAS, which typically utilize array data of mainly common markers, most often using the same or similar arrays to those we have used for our analyses (Affymetrix Human SNP array 6.0 and Illumina Human1M). We report results for association testing of all genotyped markers located within 3 cM of the resequenced disease locus, after verifying that the vast majority of significant associations are within those bounds (Materials and Methods). Similar to the requirement of genome-wide significance in a GWAS, we required significance following multiple-hypothesis correction for the entire region tested, such that our results can be extrapolated to genome-wide studies. We repeated the association testing for 5 different disease loci (Table 2.1) and for 50 sets of random assignments of causal variants in each locus. For each of these sets, we repeated the association testing in 10 replicates, varying between them only the stochastic assignment of cases and controls, for a total of 500 association tests in each locus for each of the three scenarios of causal variants. We also considered separately both a European (CEU) and a West African (YRI) population. Because of the relatively small sample size of ~110 individuals, we simulated a larger sample using HAPGEN (Spencer, Su et al. 2009), which maintains the genetic variation observed in the original data, including patterns of LD and MAF

(Materials and Methods).

All scenarios show significant associations much more often than the false discovery rate of 5% (Table 2.S1). To determine whether "synthetic associations" due to underlying rare variants tend to be further away than "natural associations" due to underlying common variants, we considered for each association test the distance between any association and the causal variant with which it is in strongest LD (Materials and Methods). We found that the median distance, over the many hundreds of associations found across the 500 tests, is variable across the five loci and—to some extent—between the two populations (Figure 2.1). Synthetic associations tend to be much further than natural associations, as previously predicted (Dickson, Wang et al. 2010), though for one region (disease locus #1) both synthetic and natural associations are in close proximity to the causal variants (Figure 2.1). Alternatively, when considering the distance between an association and the closest causal variant (rather than the one in strongest LD), the distance of synthetic associations is reduced, yet generally remains greater than that of natural associations (Figure 2.S1). Taken together, these results lead us to ask what factors contribute to this increased distance, and, more importantly, does this increased distance impact the choice of fine-mapping strategies?

We explored several plausible explanations for this increased distance. Firstly, we ensured that the increased distance of rare causal variants is not due to more variants in those scenarios (5 and 9) than in the scenario of common causal variants (2) by repeating our analysis for cases with 5 common causal variants. We observed no increase in association distance of resultant natural associations (Figure 2.S2), revealing that the increased distance is not due to the increased

number of causal variants. Secondly, we investigated the hypothesis that increased marker effect size can cause greater association distances as effect size, in addition to the correlation between the causal variant and the marker, is proportional to the power of an association test (Chapman, Cooper et al. 2003). We investigated this hypothesis by increasing the effect size of common causal variants to equal that in the scenario of rare causal variants, though such an effect size might be considered unrealistic for common variants. The median association distance of the resulting natural associations indeed increases for all regions and populations, but is still considerably lower than synthetic associations in most cases (Figure 2.1).

We next tested whether the age of the mutation played a role in increasing association distances for synthetic associations. As rare variants are, on average, resultant of more recent mutations compared to common variants, recombination would have had less time to operate, thus resulting in diminished decay of LD and haplotype structure around rare variants. To test whether the age of the mutation plays a part in explaining our results, we partitioned rare causal variants in two age groups: i) variants due to relatively *more recent* mutations and ii) variants due to relatively *older* mutations. Variants with minor alleles present in only a single population fell into the former category, while those with minor alleles present in more than one population fell into the latter (Materials and Methods). We observed a larger distance between an associated marker and the causal variant with which it is in highest LD for *more recent* mutations than for *older* mutations (Figure 2.2). Out of the 4 disease loci for which enough data was available to perform this analysis, 3 in YRI and 2 in CEU exhibit a median distance from *older* rare causal variants that is at least 41% less than the median distance from *more recent* causal variants. Combined, these results suggest that the increased distance of synthetic associations compared to natural

associations is partially due to the young age of the mutations that give rise to rare risk alleles, as well as due to the higher effect size that is likely to be implicated for rare risk alleles.

The main concern regarding synthetic associations is how its signatures alter the search for the actual causal variant(s). Specifically, how far should one sequence around an association in order to capture causal variants? We addressed this question using two approaches. We first computed for each scenario of causal variants the fraction of tests (out of all tests with any significant association) that had at least one associated marker within any given distance of the causal variant with which it is in highest LD (Materials and Method). We found that for common causal variants, a shorter resequencing distance of 0.01 cM is enough to capture a causal variant in 90% of the tests in CEU and 77% for YRI (Figure 2.3). For rare causal variants, combined over all disease loci, at least 90% of tests discovered an association within 0.1 cM of a causal variant (Figure 2.3). Secondly, we investigated a scenario in which fine-mapping consists of sequencing the LD block of associations as observed in the data. Hence, we estimated the probability that an associated marker is in the same LD block as any of the causal variants, with the definition of LD blocks being based only on markers from the genotyping arrays, which are relatively common (Materials and Method). On average, the LD blocks spanned 0.007 cM for CEU and 0.005 cM for YRI, including the addition of a flanking region of 0.0005 cM to be inclusive. We found that in CEU, 94% of associated markers were in the same LD block as a common causal variant, while the same was true for only 78% of associated markers in the rare causal variant case. A similar trend was observed for YRI, albeit less marked, where 79% of natural associations captured a causal variant, but only 73% of synthetic associations captured a causal variant.

Finally, we explored the minor allele frequency (MAF) of associated markers. Summing over all disease loci and populations, <1% of natural associations had MAF below 0.1, while this proportion increased to 15-28% for synthetic associations (Figure 2.4). Dissecting the signal further by region and population, we found that while some regions display less than 2.4% difference between the median MAF of natural associations and synthetic associations (disease locus #1 in YRI, #2 in CEU), others display an almost 200% difference (#4 in CEU). Synthetic associations also display a larger standard deviation in associated MAF as compared to natural associations, with all but one region displaying a difference ranging from 17%- 70% (Table 2.2).

2.3 Discussion

With the use of HapMap 3 resequencing and genotyping data from five different genomic regions and two populations (Altshuler, Gibbs et al. 2010), we considered several scenarios of disease risk loci, and performed association tests to investigate the signatures of synthetic associations and how they alter one's approach for studying them. We found that the median distance of synthetic associations, while greater than that of natural associations, still never exceeds 0.15 cM (~150 kb) for any of the 10 locus-by-population settings. Even if we instead consider the worst-case scenario of the largest distance between any association and any causal variant, its median still never exceeds 0.41 cM (~410 kb). These results are in clear contrast to the results of a previous simulation-based study that showed the median of the largest distance to be 5 cM (5 Mb) (Dickson, Wang et al. 2010). The difference between the two studies may be attributed to differences in the frequencies of rare causal variants. We considered rare alleles of

frequency in the range 0.005-0.04 (average across all variants of 0.019), while Dickson et al. simulated allele frequencies in the range 0.005-0.02 (Dickson, Wang et al. 2010) (average of 0.0125 assuming uniform sampling). However, when we restricted to a narrower range of frequencies up to 0.02 (average of 0.012), we still observed no locus for which the median distance of synthetic association exceeds 0.5cM ('All variants' in Figure 2.2). It is unlikely that any remaining slight difference in risk allele frequency would result in over an order of magnitude difference in association distance.

A more substantial difference between the two studies lies in the data analyzed. Dickson et al. conducted simulations of constant effective population size, uniform recombination rate, and purely neutral disease loci, with association testing based on a simulated "genotyping array" that follows a uniform ascertainment bias (Dickson, Wang et al. 2010). Here, we have analyzed data with empirically observed LD patterns, and have based association testing on data from real genotyping arrays as designed for GWAS. Put together, while theory posits that a median distance of synthetic associations of 5cM is possible, characteristics of empirical data suggests that such cases will not be common, and that even under the worst-case scenario the vast majority of synthetic associations are at least an order of magnitude closer.

By considering which of the rare polymorphisms are population-specific, and hence likely to be more recent, we illustrated that the increase in association distance can partially be due to the age of the mutation. This is likely a result of recombination having had less time to break down the haplotype surrounding more recent mutations. We also considered common causal variants with a higher effect size and showed that an increased effect size can lead to an increased association

distance. As rare causal variants contributing to an association signal are likely to have higher effect sizes than common causal variants, the increased distance for synthetic associations can thus partially be due to the larger effect size. Our findings thus suggest that synthetic associations do not necessarily entail further causal variants. While these explanations apply in scenarios where single causal variants contribute to an association signal, the increased distance for synthetic associations can also result from the contribution of multiple causal variants to a single signal of association, thus exceeding the expectations of distance given two variants in LD.

To assess the impact of this increased association distance, we explored the probability that an association test had at least one association where the causal variant with which it was in highest LD lay within a given distance from the association. We found that for rare causal variants a window size of 0.1 cM was sufficient to capture at least one causal variant in such a manner in at least 90% of the tests for all regions and populations (Figure 2.3). Alternatively, by following an LD block based approach for fine-mapping, 73-79% of associations capture at least one of the rare causal variants within the same LD block. This suggests that traditional LD block-based fine-mapping also offers a surprisingly high probability of discovering some of the causal variants, though there could still be added benefit from sequencing a larger region. Thus, given a resequenced region, one is almost as likely to capture a rare causal variant as one is to capture a common causal variant. Preliminary analysis suggests that it is difficult to predict the optimal region to resequence given a specific disease locus, as no single factor such (i.e. pair-wise LD decay) can sufficiently predict this distance (data not shown). Further work is thus necessary in order to determine which factors that influence synthetic associations, such as the age of mutation, causal variant effect size, haplotype structure and the stochastic coupling of multiple

rare variants on the background of a common marker, play a role in a given association signal.

In a further analysis, we found that rare causal variants underlying synthetic associations entail that the associated markers will themselves be of lower frequency compared with natural associations (Figure 2.4), a result consistent with previous simulation studies (Dickson, Wang et al. 2010; Wray, Purcell et al. 2011). When narrowing the number of associations to only the most significant, we found that this further reduced the allele frequency of synthetic associations (Figure 2.S3). In addition, we found that the frequency of synthetic associations often had a larger standard deviation than natural associations (Table 2.2). These results have two implications. Firstly, it suggests that synthetic associations as compared to natural associations are likely to have underestimated effect sizes of the causal variant due to reduced associated allele frequencies (Spencer, Hechter et al. 2011) (especially when analyzing the most significant association) and from incomplete LD with the causal variant. Secondly, this suggests that the standard deviation of the associated minor allele frequency can offer a way to flag for underlying rare causal variants that induce potential synthetic associations; given a larger standard deviation of associated frequencies, it would be advised to follow a fine-mapping study design for synthetic associations.

With the >1000-fold human population growth in the last hundreds of generations, the amount of rare variation is much greater than expected (Coventry, Bull-Otterson et al. 2010; Keinan and Clark 2012). This explosive addition of rare variation entails an LD structure that is yet to be quantified, but certainly disparate than the LD structure of common variants that have been extensively studied. In addition, the earlier founder events as modern humans migrated out of

Africa and settled across the globe have been shown to greatly alter patterns of genetic variation (Tishkoff, Dietzsch et al. 1996; Dunning, Durocher et al. 2000; Reich, Cargill et al. 2001; Ramachandran, Deshpande et al. 2005; Keinan, Mullikin et al. 2007). For this reason we studied both a West African population and a population of European ancestry, with differences between the two reinforcing the importance of taking demographic history into consideration by studying empirical data. Signatures of synthetic and natural associations are shaped by demographic history, as well as by different selective pressures. This assertion is supported by the highly variable behavior—across genomic regions and across the two populations—of all the signatures we observed.

In conclusion, this study delivered a characterization of several signatures of synthetic associations and assessed their impact on the search for the causal variant(s) underlying the signal. While our study does not participate in the debate on how frequently synthetic associations occur, it is relevant in any situation they do. In this study, we illustrated that because synthetic associations are likely to be more distant from causal variants, fine-mapping studies should look further than when searching for common causal variants, but to a much lesser extent than previously suggested. We also propose the larger standard deviation of associated allele frequencies as a way to detect potential rare causal variants at play. Additional analysis is warranted though, to elucidate the quantitative relationship between genetic architecture, demographic history, allele frequency and association signals. Finally, although the debate remains open as to the contribution of rare risk alleles to human complex diseases and to the ensuing abundance of synthetic associations (Orozco, Barrett et al. 2010; Anderson, Soranzo et al. 2011; Goldstein 2011; Wray, Purcell et al. 2011), our results offer new guiding principles for

determining a length of a region to fine map, and for considering the alternative of an association study based on whole-genome sequencing.

2.4 Materials and Methods

2.4.1 Data

We obtained from HapMap 3 (Altshuler, Gibbs et al. 2010) genotyping array data for YRI (Yoruba in Ibadan, Nigeria) and CEU (individuals in Utah with Northern and Western European ancestry from the Centre d'Etude du Polymorphisme Humain collection) and resequencing data of five ENCODE regions, each 100kb in length (Table 2.1), for 115 YRI and 111 CEU individuals. We also obtained resequencing data for 60 TSI (Toscani in Italia) samples and 60 LWK (Luhya in Webuye, Kenya), which we used for the *variant age analysis* (below). We considered each resequencing region as a disease locus from which to select causal variants. Using resequencing data facilitates higher concentration of rare variants and is free of the ascertainment biases associated with genotyping arrays (Clark, Hubisz et al. 2005; Frazer, Ballinger et al. 2007; Keinan, Mullikin et al. 2007; Albrechtsen, Nielsen et al. 2010).

2.4.2 Simulated Data

Due to the low sample size, we employed HAPGEN (Spencer, Su et al. 2009) to simulate 10,000 individuals for each population –a strategy previously employed to investigate the estimation of relative risks (Spencer, Hechter et al. 2011). HAPGEN simulates additional haplotypes by treating each new haplotype as a mosaic of already present haplotypes. We refer readers to (Spencer, Su et al. 2009) for additional details on HAPGEN.

We first phased and imputed missing data with BEAGLE v3.3 (Browning and Browning 2007). We then simulated additional data for each resequencing region and the 3 cM-flanking window for each region using HAPGEN with a recombination map from the March 2006 human reference sequence (NCBI Build 36, hg18) and a null mutation rate as input parameters. We ensured that the LD patterns of the original data (for rare and common variants) were maintained (Figure 2.S4). We also ensured that allele frequencies in the simulated data do not change drastically from the original data as no variants were observed that were initially of very low frequency and attained a much higher frequency and vice versa in the simulated dataset (Figure 2.S5-2.S6).

Association tests were performed using the simulated data from the HapMap 3 genotyping array data, excluding any causal variants that happen to be in the genotyping array data. We report results for an association study for SNPs located in the disease locus and in flanking regions of 3 cM on each side (from which no causal variants are chosen), as almost no associations were observed to fall beyond that distance (data not shown). In our study, rare causal variants have risk allele frequencies in the simulated data between 0.005 and 0.04 (we note that a portion of this range is defined as "low frequency", rather than rare, by some studies), and common causal variants have risk allele frequencies in the simulated data between 0.1 and 0.3. In testing for association, we considered all SNPs of all allele frequencies from the genotyping data. All coordinates and genetic distances in this paper are according to the March 2006 human reference sequence (NCBI Build 36, hg18).

2.4.3 Disease model & association study design

We considered each individual as a case or a control with a probability proportional to the individual's assigned risk, which is elevated if the individual has one or more risk alleles. We set the baseline risk as 0.15 and the genotypic relative risk to 1.5 for the scenario of common causal variants. We also explored an unrealistic genotypic relative risk of 3 for common causal variants to investigate the influence of effect size on association distance. For rare causal variants, we assigned a higher genotypic relative risk of 3. While the use of a fixed GRR for variants of differing allele frequencies results in differing portions of variance explained by each variant, it is a more realistic disease model. By fixing variance explained, rarer variants would tend to have higher, and perhaps somewhat unrealistic, GRRs. Because we have fixed GRR and allowed the proportion of variance explained to vary, an association test will have more power in detecting variants of higher allele frequency given a fixed GRR.

For the common causal variants scenario, we randomly assigned 2 SNPs from the resequencing data as causal, while we assigned either 5 or 9 for the rare causal variants scenario. To ensure that the number of causal variants did not affect our results, we also studied a scenario with 5 common causal variants in loci where this was feasible. For each scenario of a certain type and number of causal variants, 50 sets of causal variants were randomly selected, with replacement between groups. Each of these 50 sets allows for a possibly different risk for each individual. For each of these 50 sets, we repeated 10 replicates of randomly assigning cases and controls according to the same individual assigned risk.

In each of the 500 association tests (50 different variant groups and their 10 phenotypic

replicates), we randomly chose 1000 cases and 1000 controls according to the individual's assigned risk. This ensures that the same number of cases and controls were shared across all analyses, thereby having comparable statistical power. For each scenario of type and number of causal variants, we pooled together the results from these 500 tests for the statistics and figures presented in this study. Similarly, we generated 500 tests for each disease locus with randomly assigned case/control status to serve as a control.

All association tests were done with PLINK's logistic regression function (Purcell, Neale et al. 2007). Significance thresholds were determined with a region-wide Bonferroni correction. For the control scenario of random assignment of cases and controls, 2.12% of the association tests showed a significant association as compared with the expectation of 5%.

2.4.4 Distance analysis

We determined genetic distances based on the Oxford genetic map based on HapMap2 data (Myers, Bottolo et al. 2005; Frazer, Ballinger et al. 2007). For SNPs missing from HapMap2, we estimated the position as the linear interpolation of the genetic positions of the two closest SNPs. The association distances were determined by computing the genetic distance between an associated SNP and the causal variant with which it was in highest LD, measured in r^2. Pairwise r^2 values were calculated in pLINK (Purcell, Neale et al. 2007).

2.4.5 Age of mutation analysis

To partition rare variants based on the age of the mutation, we first narrowed the range of the risk allele frequency in the simulated data to 0.005 and 0.02 in order to ensure a roughly equal partition into the two age groups. We discarded disease locus #1 from this analysis because it had too few rare variants to allow their portioning into two groups (Table 2.1). Rare variants in the 111 CEU individuals were defined to be relatively *more recent* if only the major allele was observed in the resequencing of 115 YRI individuals and 60 TSI individuals in the original data; the variant was defined as relatively *older* otherwise. We repeated the above analyses for each of these groups separately, such that in each association testing either all causal variants are *older* or all are *more recent*. We repeated the same analysis in YRI with CEU and 60 LWK as out groups. We duly note that polymorphisms absent from the limited number of samples may not be monomorphic in the population as a whole, hence not all mutations leading to relatively *older* variants precede those leading to variants in the relatively *more recent* class. Yet, this represents only a small fraction of variants and variants in the relatively *older* class are expected to be older on average than those belonging to the *more recent* class. It is also important to note that false positive variant calls are added to the *more recent* group despite the erroneous call. This scenario is highly unlikely in our analyses due to the stringent quality control measures taken in HapMap 3 [45] and the exclusion of singletons in our study. For each of these two scenarios of causal variants, we similarly chose 50 sets of causal variant groups with 10 phenotypic replicates each and obtained maximal distances as above. For comparison, we repeated the analysis for random rare causal variants in the narrowed range of frequency of 0.005-0.02 used here, irrespective of mutation age.

2.4.6 Resequencing distance analysis

For each association test we explored whether a causal variant with which an association is in highest LD (measured in r^2) is within a given genetic distance from the association. For each simulated scenario and resequencing window size ranging from 0 cM to 10 cM, we calculated the proportion of tests that have at least one such association.

For the second analysis, we observed over all significant associations if any causal variant was in the same LD block as an association. LD blocks were estimated in pLINK with the genotyping data (Purcell, Neale et al. 2007) and 0.0005 cM was added to the start and end coordinates in order to compensate for the uncertainty in these estimates.

2.5 Figures

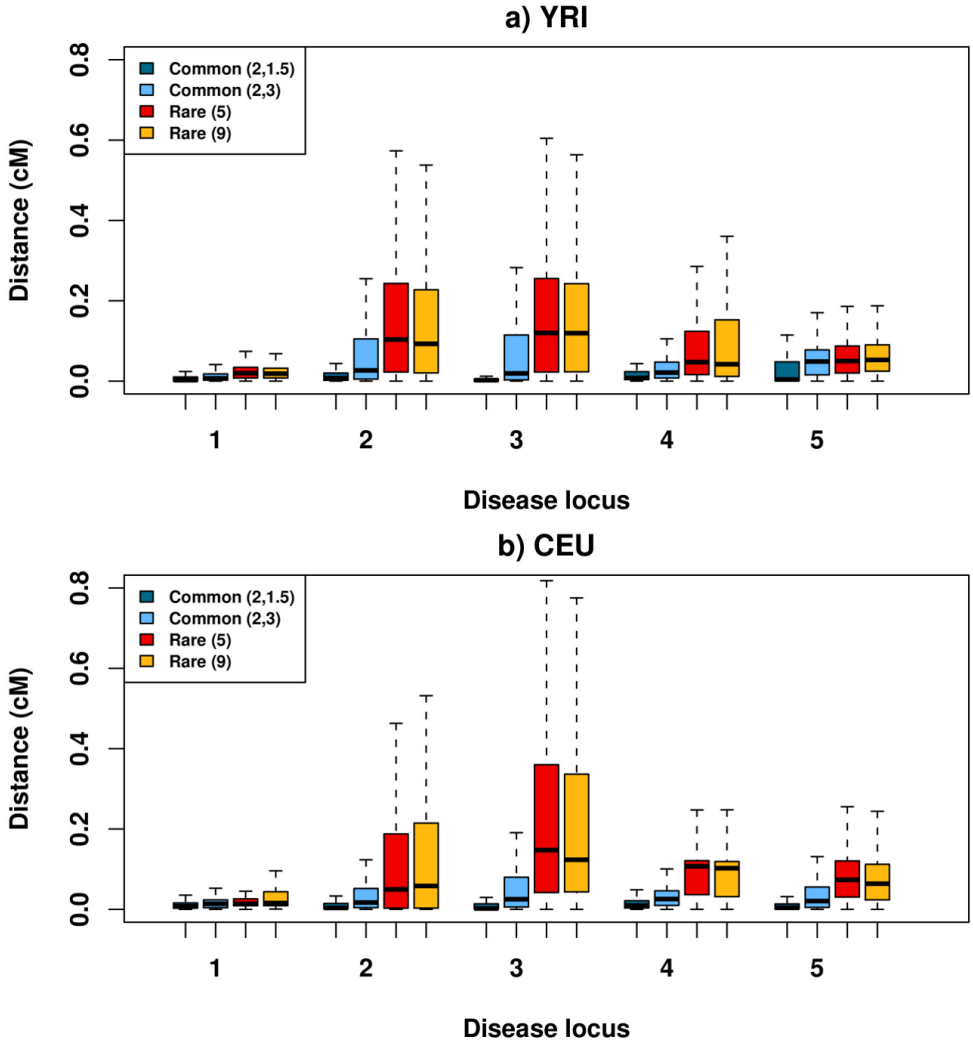

Figure 2.1. Distance of synthetic and natural associations from the causal variant it is in greatest LD with.
Box plot of the distance between any associated SNP and causal variant it is in highest LD with, measured in r^2, for **(a)** YRI and **(b)** CEU in four scenarios: 2 common causal variants with a GRR of 1.5 (dark blue), 2 common causal variants with an unrealistic GRR of 3 (light blue), 5 and 9 rare causal variants with a GRR of 3 (red and gold respectively). Distances vary greatly between the different disease loci (x-axis) as well as between populations, but in all regions the median (line within each box) is larger for rare causal variants than for common causal variants of lower effect size. Increasing the effect size can result in higher association distance as is observed most notably in region #5.

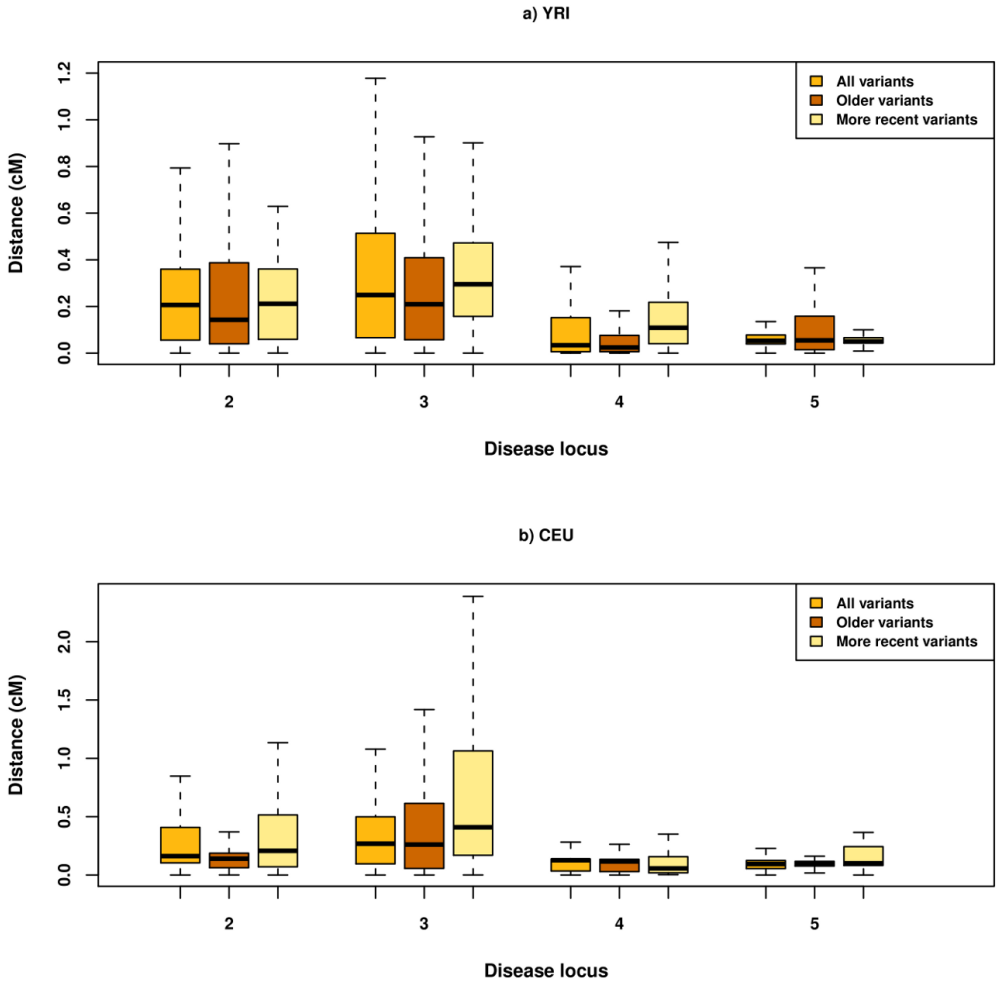

Figure 2.2: Distance of causal variant from "synthetic associations" partitioned by the age of the mutation.
Box plot similar to Figure 2.1, while separating rare variants in CEU and YRI into a *more recent* and an *older* class (Materials and Methods). Variants due to more recent mutations result in much increased distance between the associated SNP and the causal variant with highest LD in 3 regions in YRI and 2 regions in CEU. Results are presented for only 4 of the disease loci due to lack of relevant data in locus #1. Note that the risk allele frequency range for rare variants is narrower compared to Figure 2.2 (Materials and Methods) and that the y-axis scale is different between the two populations.

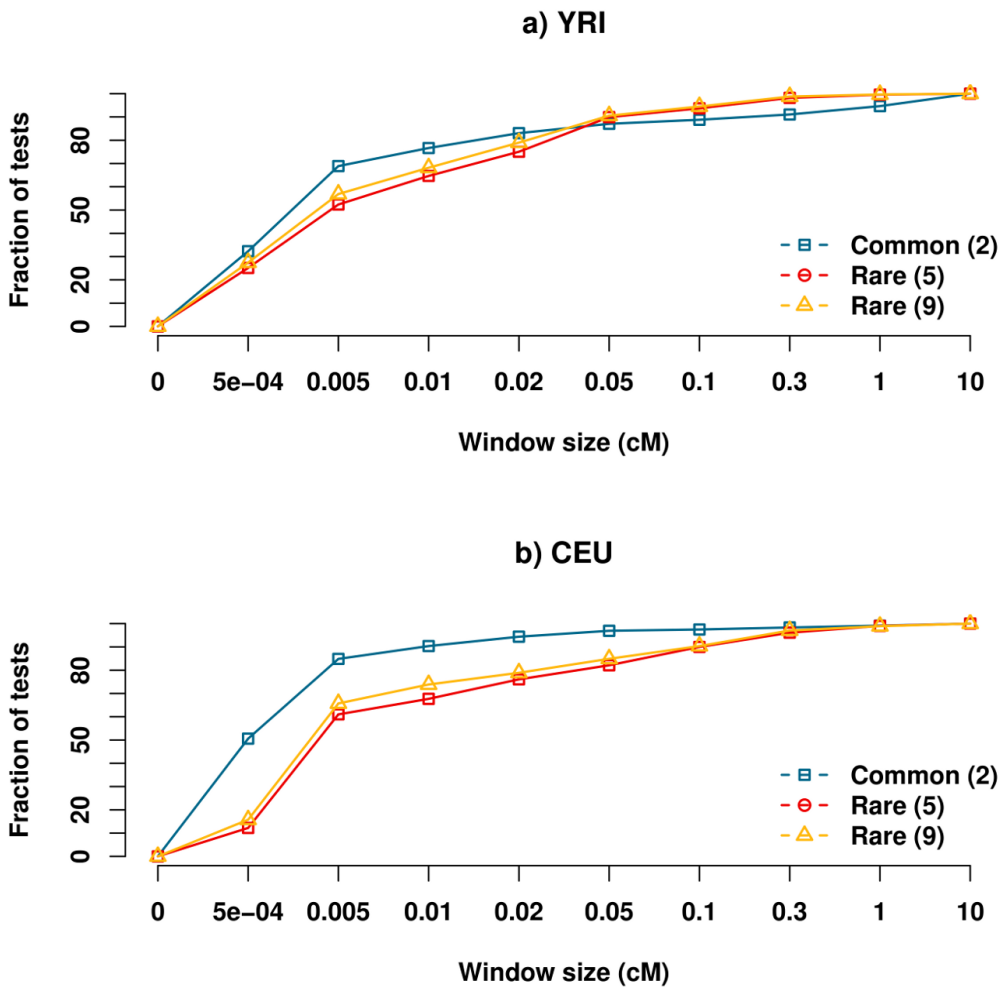

Figure 2.3: Resequence window size necessary to capture at least one causal variant. The figure presents for a given window size, the fraction of tests combined over all regions with significant associations where at least one association is within the given distance from the causal variant it is in highest LD with. The colors correspond to the same scenarios as in Figure 2.1. Resequencing need not extend much further than in the common causal variant case, as a window of size of 0.1 cM has at least one association tagging a rare causal variant in >90% of the tests between both populations and all regions.

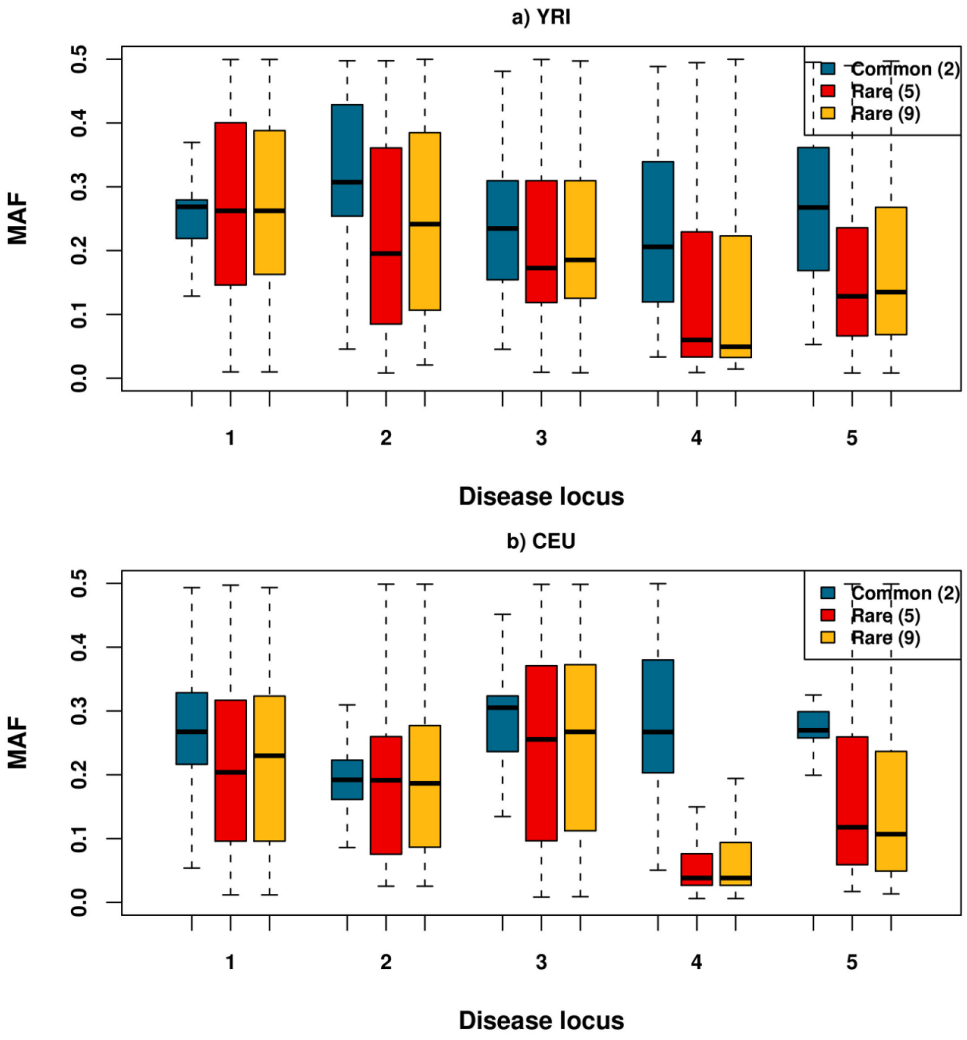

Figure 2.4. Minor allele frequency of associated variants.
Box plot of the minor allele frequency for all associated variants in the different scenarios. Although synthetic associations have median MAF lower than that of natural associations, the range of MAF for synthetic associations varies across the different loci and populations. The median MAF is similar between the natural and synthetic associations for a few loci (disease locus #2 in CEU and #1 in YRI).

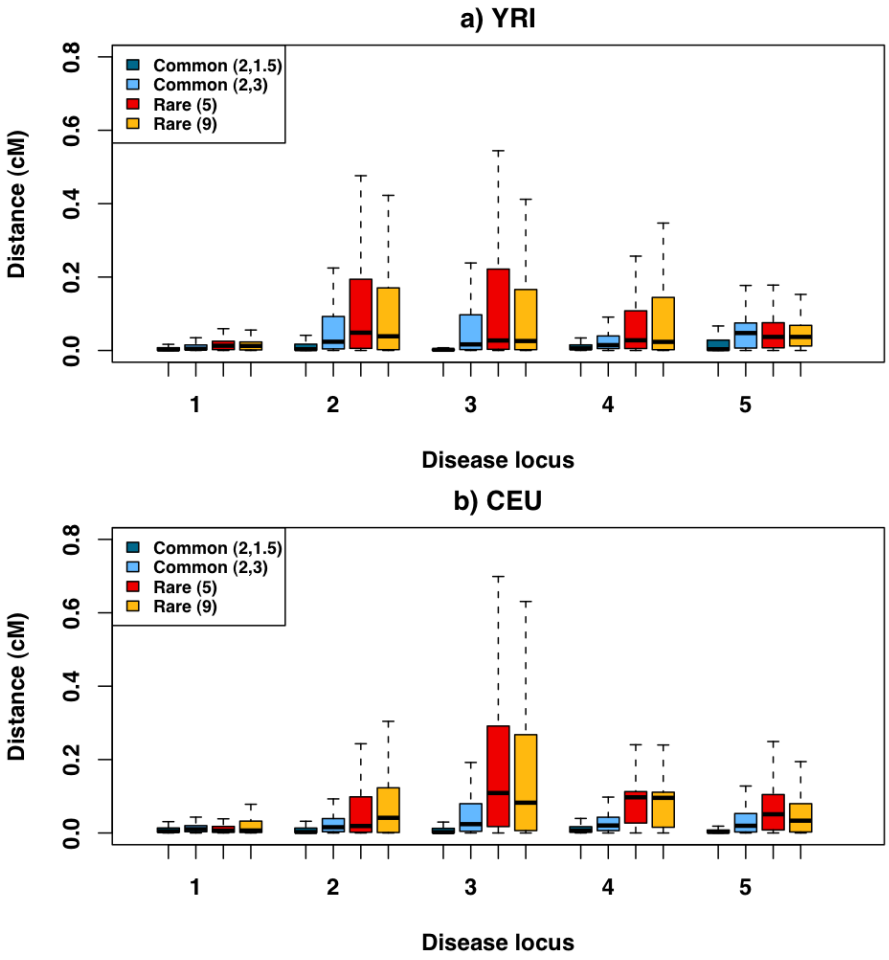

Figure 2.S1: Distance between association and closest causal variant.
The figure mirrors Figure 2.1, but plots instead the distance between an association and the closest causal variant. The distance of synthetic associations is reduced, yet generally remains greater than that of natural associations.

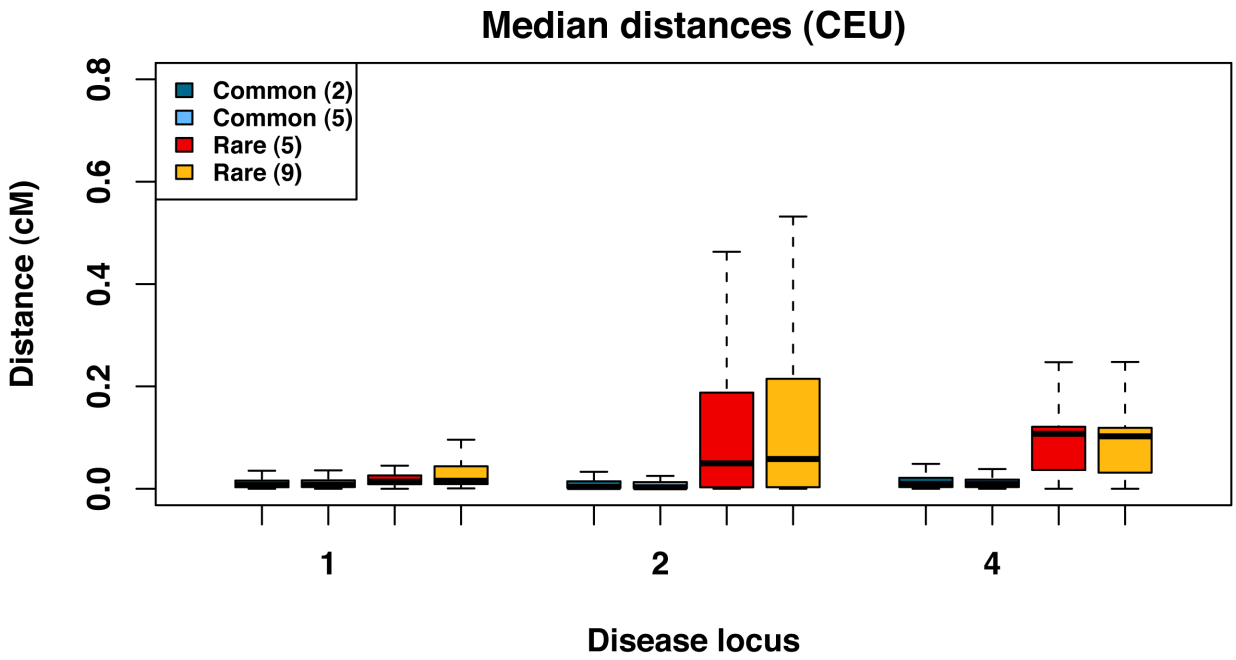

Figure 2.S2: Distance of common causal variant is not sensitive to the number of causal variants.

The figure mirrors Figure 2.1, but to the inclusion of results for 5 common causal variants ("Common (5)") in loci where this was feasible (all for CEU). All other results are reproduced from Figure 2.1. The difference in distance between common and rare causal variants remains even with 5 common causal variants.

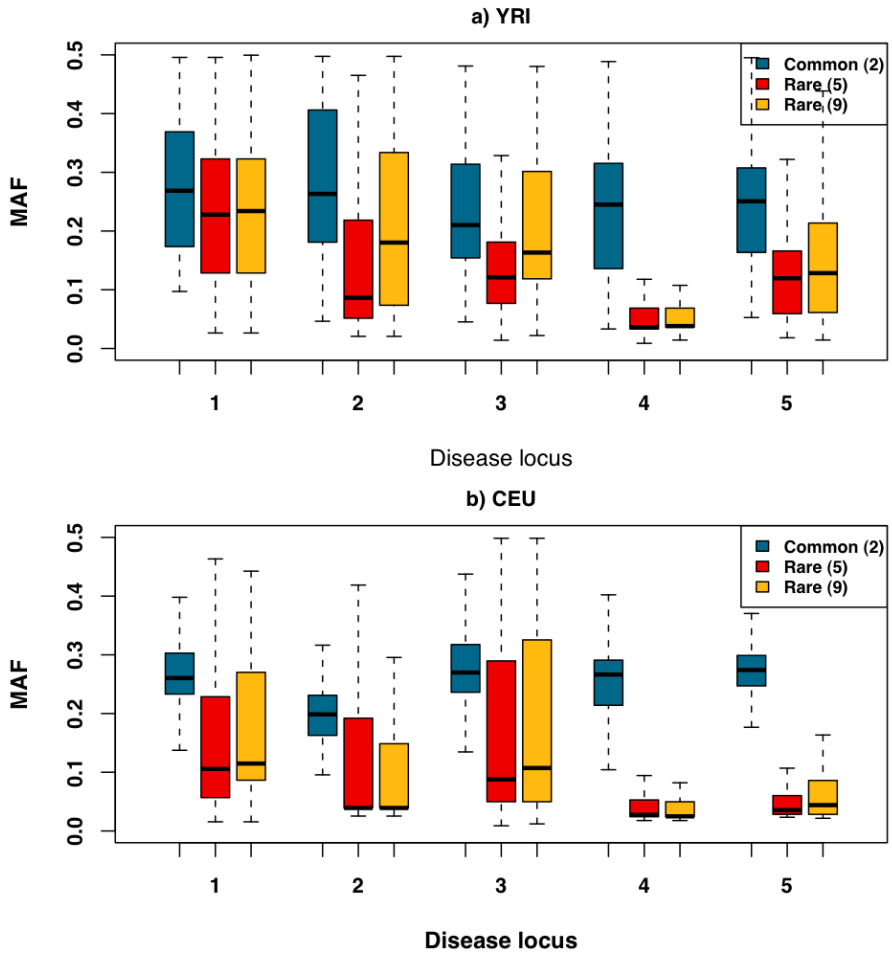

Figure 2.S3: Minor allele frequency of most significant association.
The figure mirrors Figure 2.4, but displays the minor allele frequency of only the most significant association across each test. The median frequency of the most significant association is reduced for synthetic associations.

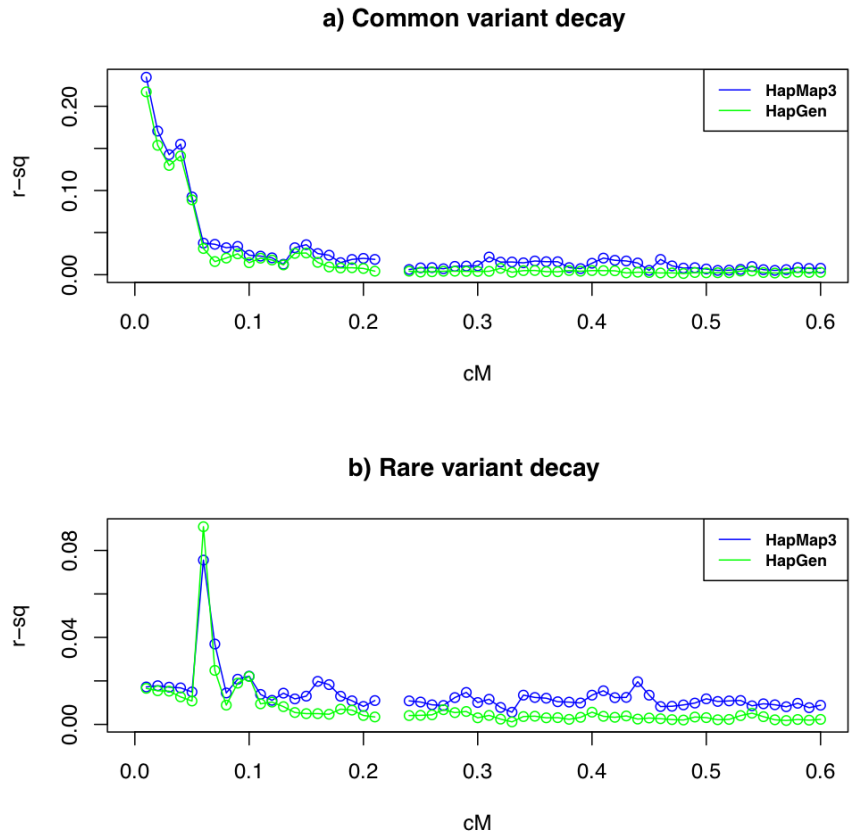

Figure 2.S4: Empirical LD patterns are preserved in HAPGEN simulations.
Plotted above is data for region 1 in CEU. For each 0.01 cM bin, the figure presents the mean pair-wise LD (measured in r^2) between variants from the resequencing and genotyping data for a) common markers (minor allele frequency > 0.04) or b) common and rare markers (minor allele frequency < 0.04). We observe that HapMap 3 LD patterns (blue) are largely preserved in HAPGEN simulations (green). Missing points reflect lack of data for certain distance bins.

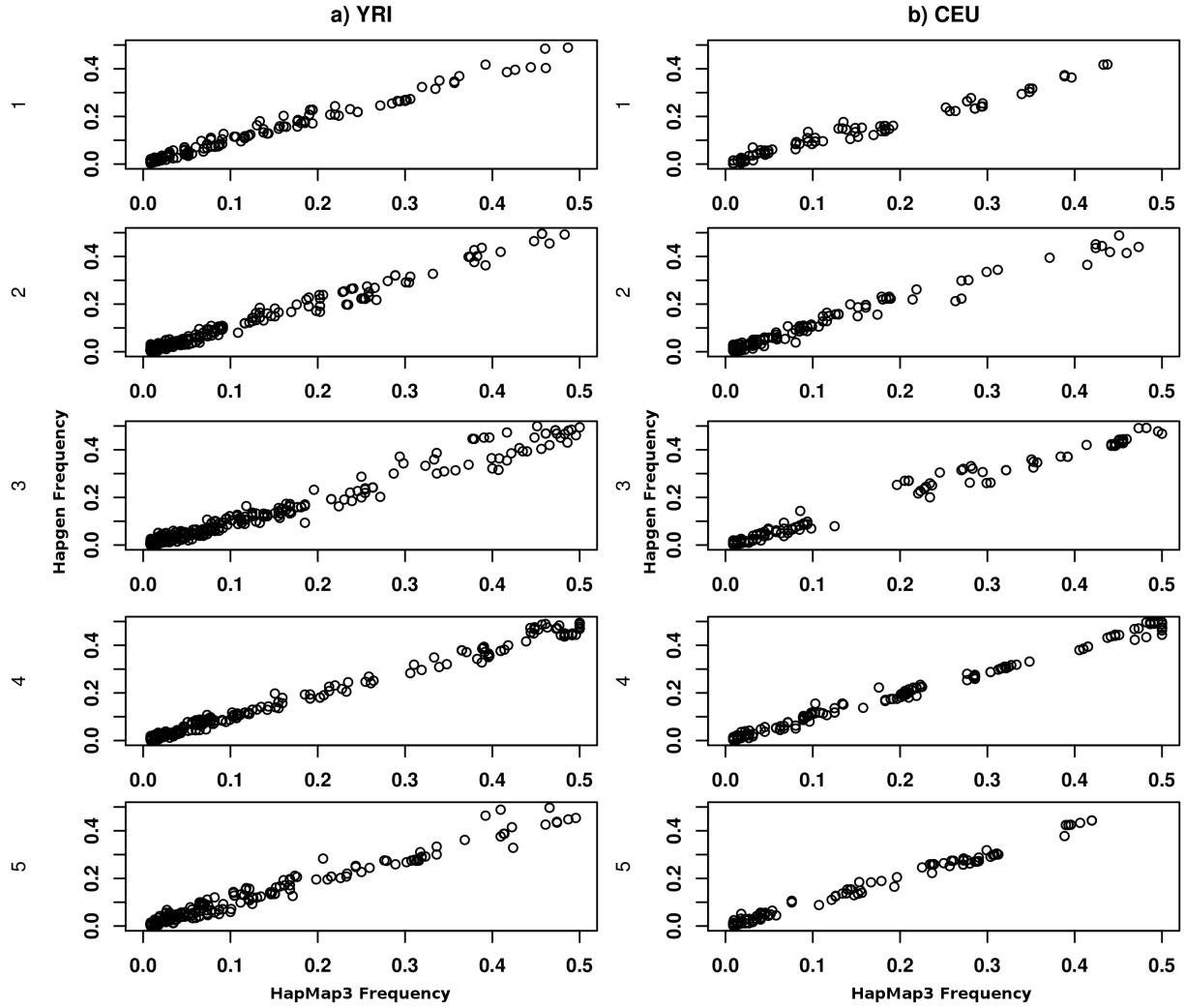

Figure 2.S5: Minor allele frequency in HapMap3 compared to minor allele frequency in HAPGEN simulations.

Plotted are minor allele frequencies in HapMap 3 (x-axis) compared to minor allele frequencies in HAPGEN simulations (y-axis) for a) YRI and b) CEU. Each row represents a separate region. No drastic departures from the original minor allele frequencies are observed in the simulated data.

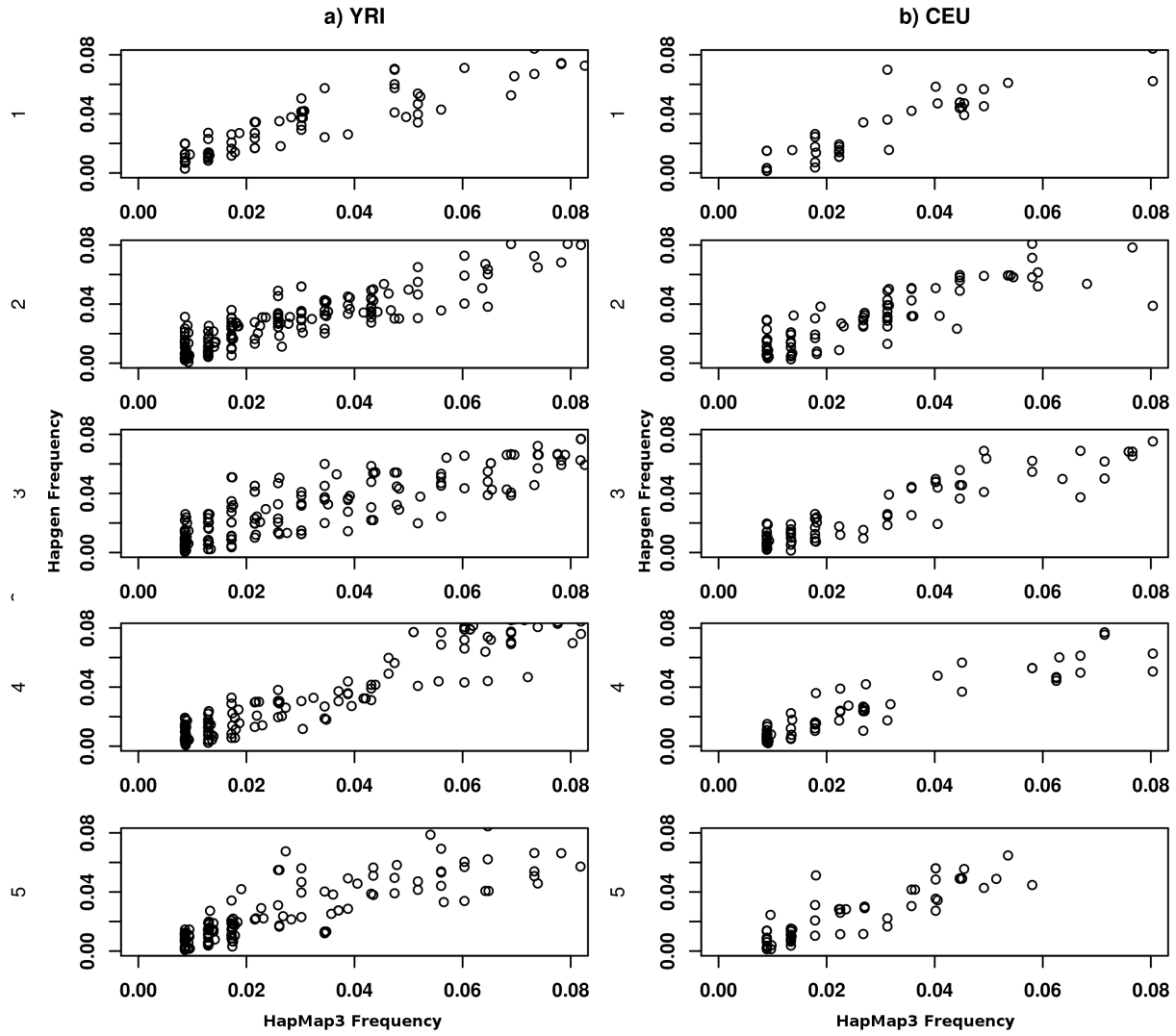

Figure 2.S6: Minor allele frequency in HapMap3 compared to minor allele frequency in HAPGEN simulations for frequencies below 0.08.
Same plot as in Figure 2.S5 showing only variants with frequencies below 0.08. As in Figure 2.S5, no drastic departures from the original minor allele frequencies are observed in the simulated data

2.6. Tables

Table 2.1: List of ENCODE regions used as disease loci [45].

Locus #	ENCODE name	Chromosome	Location (bp)	# Common variants* (YRI/CEU)	# Rare variants* (YRI/CEU)
1	ENr221	5	56071684 -56170943	57/36	59/20
2	ENm010	7	27124056 -27223436	58/40	117/57
3	ENr321	8	119082399 -119182123	72/20	108/45
4	ENr123	12	38827200 -38925373	43/62	72/50
5	ENr213	18	23920590 -24019175	60/54	108/41

*Variants with MAF of either between 0.1 - 0.3 or between 0.005 – 0.04 after resampling of haplotypes using HAPGEN.

Table 2.2: Standard deviation of minor allele frequency for associated variants.

Locus #	Common (2)	Rare (5)	Rare (9)
YRI			
1	0.086	0.134	0.131
2	0.117	0.154	0.150
3	0.114	0.131	0.124
4	0.124	0.151	0.145
5	0.113	0.121	0.126
CEU			
1	0.084	0.113	0.116
2	0.056	0.118	0.121
3	0.064	0.152	0.143
4	0.121	0.121	0.126
5	0.073	0.133	0.136

Table 2.S1: Percentage of tests with significant associations.

	YRI				
Locus #	Common (2,1.5)^	Common(2,3)^	Rare (5)^	Rare (9)^	Random*
1	31.4	100	93.8	97.2	2
2	19.8	99.8	98	97.6	2.8
3	38	100	95.8	99.6	3.2
4	31.8	99.8	86.4	94.4	2
5	28.6	100	94	97.2	1.8
Mean	**29.92**	**99.92**	**93.6**	**97.2**	**2.36**
	CEU				
Locus #	Common (2,1.5)^	Common(2,3)^	Rare (5)^	Rare (9)^	Random*
1	67.6	100	91.6	95.6	2.2
2	58	100	95	99.6	2.4
3	37.6	100	80	84.6	1.2
4	54.2	100	89.6	96.8	2.2
5	36	100	87.6	97	1.4
Mean	**50.68**	**100**	**88.76**	**94.72**	**1.88**

^ Corresponds to the notation of Figure 2.1.
* Corresponds to random phenotypic assignment.

Chapter 3: Accounting for eXentricities: Analysis of the X chromosome in GWAS reveals X-linked genes implicated in autoimmune diseases

(Chang, Gao et al. 2014)

3.1 Introduction

Over the past decade, genome-wide association studies (GWAS) have contributed to our understanding of the genetic basis of complex human disease. The role of the X chromosome (X) in such diseases remains largely unknown because the vast majority of GWAS have omitted or incorrectly analyzed X-linked data (Wise, Gyi et al. 2013). As a consequence, though X constitutes 5% of the nuclear genome and underlies almost 10% of Mendelian disorders (Hamosh, Scott et al. 2002; Hamosh, Scott et al. 2005; Amberger, Bocchini et al. 2009; Amberger, Bocchini et al. 2011), it harbors only 15 out of the 2,800 (0.5%) total significant associations for nearly 300 traits (Green and Guyer 2011; Hindorff, MacArthur et al. 2013; Wise, Gyi et al. 2013). Even this 0.5% of associations contains a higher proportion of false positives than autosomal associations, as indicated by the occurrence of fewer X-linked than autosomal associations in putatively functional loci (<40%) (Hindorff, Sethupathy et al. 2009; Green and Guyer 2011). This phenomenon is likely due to the application of tools designed for the autosomes to X. We hypothesize that X explains a portion of "missing heritability" (Maher 2008; Manolio, Collins et al. 2009), especially for the many complex human diseases that exhibit gender disparity in risk, age of onset, or symptoms. This hypothesis is motivated by the importance of X in sexually dimorphic traits in both model organisms and human Mendelian disorders. The complex human diseases most extensively studied in GWAS are highly sexually dimorphic, including autoimmune diseases (Schuurs and Verheul 1990; Beeson 1994; Chataway, Feakes et al. 1998; Whitacre, Reingold et al. 1999; Bellamy, Beyers et al. 2000; Whitacre 2001;

Lockshin 2006; Fish 2008; Sawalha, Webb et al. 2008; Shen, Fu et al. 2010; Selmi, Brunetta et al. 2012), neurological and psychiatric disorders (Harper 1984; Gater, Tansella et al. 1998; Andersen, Launer et al. 1999; Lai, Kammann et al. 1999; Goldstein, Seidman et al. 2001; Aleman, Kahn et al. 2003; Wooten, Currie et al. 2004; Pike, Carroll et al. 2009; Jazin and Cahill 2010; Baron-Cohen, Lombardo et al. 2011), cardiovascular disease (Lerner and Kannel 1986; Anderson, Odell et al. 1991; Mendelsohn and Karas 2005; Choi and McLaughlin 2007; Teslovich, Musunuru et al. 2010), and cancer (Muscat, Richie et al. 1996; Zang and Wynder 1996; Matanoski, Tao et al. 2006; Naugler, Sakurai et al. 2007). Several mechanisms underlying sexual dimorphism have been suggested (Nelson and Ostensen 1997; Confavreux, Hutchinson et al. 1998; Whitacre 2001; Ellegren and Parsch 2007; Patsopoulos, Tatsioni et al. 2007; Fish 2008; Ober, Loisel et al. 2008), including the contribution of the X chromosome (Carrel and Willard 2005; Ropers and Hamel 2005; Ross, Grafham et al. 2005; Ober, Loisel et al. 2008; Tarpey, Smith et al. 2009). Variants on chromosome X may also be more likely to show sexually dimorphic traits as compared to the autosomes. Moreover, characterizing the role of X in complex diseases can provide insight into etiological differences between males and females, as well as a unique biological perspective on disease etiology because X carries a set of genes with unique functions (Saifi and Chandra 1999; Kemkemer, Kohn et al. 2009).

X-specific problems that should be taken into consideration include, but are not limited to: 1) correlation between X-linked genotype calling error rate and the sex composition of a plate, which can lead to plate effects that correlate with sex and, hence, with any sexually dimorphic trait; 2) X-linked variants being more likely to exhibit different effects between males and females (Dobyns, Filauro et al. 2004), suggesting enhanced power of sex-stratified statistical

tests; 3) power of the analyses being affected by the smaller allelic sample size, the reduced diversity on X and other unique population genetic patterns (Keinan, Mullikin et al. 2007; Hammer, Mendez et al. 2008; Keinan, Mullikin et al. 2009; Hammer, Woerner et al. 2010; Keinan and Reich 2010; Lohmueller, Degenhardt et al. 2010; Gottipati, Arbiza et al. 2011); 4) quality control (QC) criteria that account for sex information to prevent filtering the entirety or a large fraction of the chromosome (Wise, Gyi et al. 2013); 5) sex-specific population structure leading to differential effects of population stratification (which could inflate the type I error rate (Patterson, Price et al. 2006; Price, Patterson et al. 2006; Novembre, Johnson et al. 2008)) between X and the autosomes; and 6) application of association tests designed for the autosomes, which leads to statistical inaccuracy.

In this study, we take into account several of the above problems and apply X-aware strategies to investigate the role of X in complex diseases. Recent advancements of association test statistics for X have been made (Purcell, Neale et al. 2007; Zheng, Joo et al. 2007; Clayton 2008; Clayton 2009; Thornton, Zhang et al. 2012; Tukiainen, Pirinen et al. 2014), with one study discovering new loci associated to height and fasting insulin (Tukiainen, Pirinen et al. 2014). These improvements account for some of the aforementioned problems, but are not extensively applied, and have never been applied in the context of gene-based tests of association. Here, we demonstrate that unutilized X data from hundreds of studies can be re-analyzed to uncover X-linked disease etiology. We introduce methods and software for carrying out XWAS, which include X-specific QC, imputation, association methods, tests of sex-specific effects, and gene-based tests (Materials and Methods). Though variants displaying dominance are readily exposed in males, overall hemizygosity in males reduces the effective sample size for X. We thus

increase statistical power by focusing on whole genes as functional units and combining tests of individual SNPs into gene-based tests (Neale and Sham 2004; Jorgenson and Witte 2006; Beyene, Tritchler et al. 2009; Liu, McRae et al. 2010; Li, Gui et al. 2011). This approach also surmounts issues of replication across studies with different sets of SNPs that arise from differing genotyping arrays and quality control filtering (Neale and Sham 2004; Beyene, Tritchler et al. 2009).

A promising case study for investigating the role of X in disease risk involves autoimmune diseases (AID) and other diseases with a potential autoimmune component. Most AID are sexually dimorphic with many diseases more prevalent in one sex than the other (most in females) (Whitacre, Reingold et al. 1999; Whitacre 2001; Lockshin 2006; Gleicher and Barad 2007). Furthermore, they often show sex-specific symptoms, age of onset, and progression (Schuurs and Verheul 1990; Beeson 1994; Chataway, Feakes et al. 1998; Whitacre, Reingold et al. 1999; Bellamy, Beyers et al. 2000; Whitacre 2001; Lockshin 2006; Fish 2008; Sawalha, Webb et al. 2008; Shen, Fu et al. 2010; Selmi, Brunetta et al. 2012). While pregnancy (Nelson and Ostensen 1997; Confavreux, Hutchinson et al. 1998; Whitacre 2001) and other environmental factors (Tiniakou, Costenbader et al. 2013), as well as sex hormones (Nelson and Ostensen 1997; Confavreux, Hutchinson et al. 1998; Whitacre 2001; Fish 2008), can contribute to sexually dimorphic characteristics, a role for X-linked genes has also been suggested (Ober, Loisel et al. 2008; Libert, Dejager et al. 2010; Bianchi, Lleo et al. 2011; Quintero, Amador-Patarroyo et al. 2012; Selmi, Brunetta et al. 2012), with many having immune-related functions. Though AID have been extensively studied by GWAS, the majority of previously discovered loci have a small effect size and the combined effect of all associated loci only explains a

fraction of heritable variation in disease susceptibility (Tysk, Lindberg et al. 1988; Sofaer 1993; Jostins, Ripke et al. 2012). Despite the dozens of GWAS in AID, few have studied the contribution of X and, to date, little evidence of its role in AD has been provided (Green and Guyer 2011; Hindorff, MacArthur et al. 2013; Wise, Gyi et al. 2013), though X-linked loci overall may contribute to the heritability for some complex diseases (Yang, Manolio et al. 2011). Hence, we applied the X-specific analytical methods and software developed as part of this study to conduct an extensive XWAS of a number of AID and other diseases with a potential autoimmune component (DPACs) (Pagani, Gonzalez et al. 2011; Itariu and Stulnig 2014), for a total of 16 different datasets (Table 3.1).

Our findings illuminate the potential importance of X in autoimmune disease, show that X-based analysis can be used to fruitfully mine existing datasets, and provide the tools and incentive for others to do the same. Additional XWAS can further elucidate the role of sex chromosomes in disease etiology, explore the role of sexual dimorphism and gender disparity in disease, and introduce gender-specific diagnosis and gender-specific treatment of complex disease.

3.2 Results and Discussion

3.2.1 Associations of individual X-linked genes with autoimmune disease risk

We assembled 16 datasets of AID and DPACs for analysis (Table 3.1). For each dataset, we first carried out QC that we developed expressly for the X chromosome (Materials and Methods), and excluded the pseudoautosomal regions (PARs). We then imputed SNPs across X based on whole-genome and whole-exome haplotype data from the 1000 Genomes Project (Materials and Methods). Of the 16 datasets, none of the original GWAS published had imputed variants in an X-specific manner, and only the Wellcome Trust Case Control Consortium 1 (WT1) datasets were analyzed with an X-aware strategy (The Wellcome Trust Case Control Consortium 2007). We applied three statistical methods to measure disease association for each SNP in each of the 16 datasets. The 16 datasets can be considered as independent as we ensured none had overlapping data (Materials and Methods). First, we utilized logistic regression as commonly applied in GWAS, where X-inactivation is accounted for by considering hemizygous males as equivalent to female homozygotes (FM_{02} test) (Materials and Methods). Second, we employed two similar sex-stratified (i.e. separately for each sex) regression analyses and combined them into a single test of association using Fisher's method ($FM_{F.comb}$ test) or Stouffer's method ($FM_{S.comb}$). $FM_{F.comb}$ accommodates the possibility of differential effect size and direction between males and females and is not affected by the allele coding in males, while $FM_{S.comb}$ takes in account both the sample size of males versus females and the direction of effect (Materials and Methods). We employed EIGENSOFT (Patterson, Price et al. 2006) to remove individuals of non-European descent and correct for potential population stratification. Following this correction, QQ (quantile-quantile) plots of the two tests across all SNPs in each dataset revealed no systematic bias though a couple studies display reduced power than expected (Supplementary

Fig. 3.S1).

We combined all SNP-level test statistics spanning individual genes to obtain gene-level test statistics (gene-based tests) in each of the 16 datasets for the FM_{02}, $FM_{F.comb}$ and $FM_{S.comb}$ tests. We considered genes by unique transcripts and—to also consider cis-regulatory elements—included a flanking 15 kilobase (kb) window on each side of the transcribed region. This test aggregates signals across all SNPs in each of these genes, while accounting for the structure of linkage disequilibrium (LD) within each gene (Liu, McRae et al. 2010). We combined SNP statistics with the truncated tail strength (Jiang, Zhang et al. 2011) and truncated product (Zaykin, Zhivotovsky et al. 2002) methods. Rather than consider only the single SNP with the strongest signal, these methods combine signals from the most significant SNPs, thus improving statistical power. This is particularly true for cases where a gene contains multiple risk alleles or where the causal SNP cannot itself be tested (Materials and Methods) (Huang, Chanda et al. 2011; Ma, Clark et al. 2013). Detailed results based on the SNP-level tests before combination into gene-based tests, are provided in Supplementary Text, Supplementary Figure 3.S2, and Supplementary Table 3.S1. We considered for replication genes with significance of $P < 10^{-3}$ as no gene was significant based upon a strict Bonferroni correction for the number of genes tested in each dataset (Table 3.1). We first attempted replication in a different dataset of the same or related disease, if such a dataset was available for our analysis (Table 3.1). Otherwise, motivated by the shared pathogenicity of different AID (Sirota, Schaub et al. 2009; Cotsapas, Voight et al. 2011; Sivakumaran, Agakov et al. 2011) (which is also supported by our following results), we attempted replication in all other datasets considered herein (Table 3.1).

We detected 54 unique genes that passed the initial criteria for discovery ($P < 10^{-3}$) in one or more of the 16 datasets, using the three types of tests, FM_{02}, $FM_{F.comb}$ and $FM_{S.comb}$. Of these, 38 genes passed the threshold in the FM_{02} test, 22 in $FM_{F.comb}$ test and 34 in the $FM_{S.comb}$ test (Supplementary Tables S2-S3), with overlap between the three sets. Of the 54 genes, we successfully replicated 5 in a different dataset of the same or related disease (Fig. 3.1a-c, Supplementary Table 3.S4) following a Bonferroni correction for the number of genes we attempted to replicate within each dataset (Supplementary Tables 3.S2-3.S3). These include 3 genes (*FOXP3, PPP1R3F* and *GAGE10*) in LD for the FM_{02} test and 3 genes (*PPP1R3F, GAGE12H* and *GAGE10*) in LD for the $FM_{S.comb}$ test that are associated with vitiligo, a common autoimmune disorder that is manifested in patches of depigmented skin due to abnormal destruction of melanocytes. All genes still successfully replicated when we repeated the gene-based analysis without the flanking region of 15 kb around each gene, though it remains unclear whether these represent independent signals or are still in LD with the same, potentially causal, variant(s). *FOXP3* has also been previously associated to vitiligo in a candidate gene approach (Birlea, Jin et al. 2011) and may be of particular interest as it is involved with leukocyte homeostasis, which includes negative regulation of T-cell-mediated immunity and regulation of leukocyte proliferation (Fontenot, Gavin et al. 2003; Tang and Bluestone 2008). Defects in the gene are also a known cause for an X-linked Mendelian autoimmunity-immunodeficiency syndrome (IPEX - immunodysregulation polyendocrinopathy enteropathy X-linked syndrome) (Bennett, Christie et al. 2001).

In CD, an inflammatory bowel disorder with inflammation in the ileum and some regions of the colon, we discovered an association of the gene *ARHGEF6* and further replicated the signal in

the Wellcome Trust Case Control Consortium 2 ulcerative colitis (WT2 UC) dataset. ARHGEF6 binds to a major surface protein of *H. pylori* (Baek, Lim et al. 2007), which is a gastric bacterium that may play a role in inflammatory bowel diseases (Luther, Dave et al. 2010; Jin, Chen et al. 2013).

Another gene, *CENPI*, has been independently associated with three diseases (amyotrophic lateral sclerosis (ALS), celiac disease, and vitiligo) (Supplementary Table 3.S5), with a combined p-value of 2.13×10^{-7} (Fisher's method). The association of *CENPI* when combining across all 16 datasets is still significant following a conservative Bonferroni correction for the number of genes we tested ($P = 2.71 \times 10^{-5}$). CENPI is a member of a protein complex that is recruited to the centromeres to participate in the assembly of kinetochore proteins, as well as generate spindle assembly checkpoint signals required for cell progression through mitosis (Matson, Demirel et al. 2012). A previous study demonstrated that it is targeted by the immune system in some scleroderma patients (Hamdouch, Rodriguez et al. 2011). Additionally, other genes in the same family have been previously associated with immune-related diseases, such as multiple sclerosis (*CENPC1*) (Baranzini, Wang et al. 2009) and ALS (*CENPV*) (Ahmeti, Ajroud-Driss et al. 2013). These findings combined, suggest a possible general role for *CENPI* in autoimmunity. Motivated by this association of *CENPI* in multiple AID and DPACs, as well as previous evidence of shared pathogenicity across different AID[87,88], we sought to replicate genes in diseases different than the ones in which they had been discovered. We successfully replicated 17 additional genes in this fashion, which we present here as suggestive evidence of these genes having a role in autoimmunity or immune-response (Fig. 3.1a-c, Supplementary Table 3.S6).

3.2.2 The sex-specific nature of X-linked genes implicated in autoimmune disease risk

If X underlies part of the sexual dimorphism in complex diseases, then we would expect some genes to have significantly different effect sizes between males and females. We tested this expectation across all SNPs and datasets (Materials and Methods). QQ-plots revealed no systematic inflation (Supplementary Fig. 3.S3). As with our analysis above, we combined these p-values to obtain gene-based tests of sex-differentiated effect size. This aims to capture a scenario whereby SNPs within the tested gene display different effects in males and females, with no constraint on a consistent direction of effect differences. We discovered and replicated *C1GALT1C1* as exhibiting sex-differentiated effect size in risk of IBD (Fig. 3.1d, Supplementary Table 3.S4; Supplementary Table 3.S7). C1GALT1C1 (also known as Cosmc) is necessary for the synthesis of many O-glycan proteins (Ju and Cummings 2005), which are components of several antigents. Defects of C1GALT1C1 may cause Tn Syndrome (a hematological disorder) (Thurnher, Clausen et al. 1992). When considering replication in datasets of other diseases, we found that both *CENPI* and *MCF2,* which we previously associated to risk of AID in our analyses above, also showed significant sex-differentiated effect sizes (Fig. 3.1d, Supplementary Table 3.S6).

We further found that some X-linked genes associated to AID exhibit differences in expression between males and females. Using a comprehensive dataset of whole blood gene expression from 881 individuals (409 males and 472 females; Materials and Methods), we assayed gene expression in males and females separately. Overall, X-linked genes that we analyzed exhibit a 2.55-fold enrichment for differential expression between males and females as compared to all genes in the human genome ($P=6.29 \times 10^{-8}$), with *XIST*, the gene responsible for X-inactivation in

females, displaying the most significant difference between males and females among all X-linked genes ($P<<10^{-16}$). Within the associated genes, four have significant sex-differential gene expression: *ITM2A* ($4.54x10^{-9}$), *EFHC2* ($4.86x10^{-5}$), *PPP1R3F* ($7.06x10^{-5}$) and *BEND2* ($4.17x10^{-4}$) (Materials and Methods). Importantly, as described above, we discovered that two of these genes (*EFHC2* and *BEND2*) exhibit sex-differentiated effect sizes, with the results herein proposing these to potentially be related to sex-differential expression patterns.

3.2.3 Biological relevance of disease risk genes

As the X chromosome carries on it a set of unique genes, we set out to explore the biological function of our associated disease risk genes. By investigating the gene expression patterns of 13 genes for which we could obtain tissue-specific expression data, (Materials and Methods), we found that three genes show the highest expression in cells and organs directly involved in the immune system (Fig. 3.3): *ARHGEF6* is expressed in T-cells, *IL13RA1* in CD14+ monocytes, and *ITM2A* in the thymus (in which T-cells develop). In addition, three other genes, *MCF2* (associated with vitiligo), *NAPL12* and *TMEM35* (associated with ALS) exhibit the highest expression levels in the pineal gland (a four-fold enrichment relative to all X-linked genes we tested, $P=3.35x10^{-3}$). The pineal gland produces and secretes melatonin, which interacts with the immune system (Calvo, Gonzalez-Yanes et al. 2013; Pohanka 2013) and has been implicated in the diseases we associated these genes to (Slominski, Paus et al. 1989; Jacob, Poeggeler et al. 2002; Sospedra and Martin 2005; Terry, Villinger et al. 2009; Dibner, Schibler et al. 2010; Calvo, Gonzalez-Yanes et al. 2013), as well as suggested as a possible treatment for ALS (Weishaupt, Bartels et al. 2006). In addition to these genes, *NLGNX4*, which is associated with psoriasis in the current study, is primarily expressed in the amygdala. Although the amygdala is

not known to affect the immune system, it mediates many physiological responses to stress (Roozendaal, McEwen et al. 2009; Mahan and Ressler 2012), which is believed to play a significant role in susceptibility to psoriasis (Heller, Lee et al. 2011).

The nature of the diseases we analyzed and the uniqueness of X led us to an *a priori* hypothesis that genes of a specific biological nature contribute to X-linked AD disease risk. Hence, we tested for association of whole gene sets with each AD or DPAC (Materials and Methods). The first two sets include X-linked genes with immune-related function as defined by the KEGG/GO or Panther databases (Materials and Methods). The third set includes the 19 non-pseudoautosomal X genes with functional Y homologs. While analysis of the immune-related gene sets was motivated by the nature of the diseases, our test of the latter set was motivated by an evolutionary perspective. X genes with functional Y homologs are more likely to be under functional constraint (Wilson Sayres and Makova 2013) and thus, may be more likely to play a part in disease etiology. We associated the Panther immunity gene set to vitiligo risk in both vitiligo studies (Vitiligo GWAS1 and GWAS2) and one type-2 diabetes study (T2D GENEVA), and the KEGG/GO set in Vitiligo GWAS1 (Table 3.2). Furthermore, genes with functional Y homologs contribute to psoriasis (CASP dataset) and vitiligo (Vitiligo GWAS1) disease risk (Table 3.2). These genes are likely to encode biologically conserved functions, as their Y homologs have retained function despite loss of recombination with X (which has led to progressive degeneration of the Y chromosome over the course of the evolution of the supercohort *Theria*) (Wilson Sayres and Makova 2013)

3.2.4 Relation between associated disease risk genes

Investigating the co-expression of associated disease-risk genes across the 881 individuals (Materials and Methods) we found that 3.9% of all X-linked gene pairs exhibit significant positively correlated gene expression patterns. Pairs of genes associated with any AID or DPAC exhibit significant positively correlated expression in 8% of the cases - a significantly higher fraction relative to X-linked genes overall ($P=1.53x10^{-3}$). This suggests that these genes are more likely to work in concert and perhaps interact in the same pathways or cellular networks. Indeed, using data from protein-protein or genetic interactions (Materials and Methods), we found that all but four are included in the same interaction network (Fig. 3.4). Perhaps not surprisingly, we found several of the significantly enriched pathways relate to immune response or specific immune-related disorders or diseases (Table 3.3). Of the remaining pathways, the regulation of actin cytoskeleton has also been found to influence the developing morphology and movement of T-cells, while the TGF-beta signaling pathway and the ECF-receptor interaction pathway can both mediate apoptosis (Lukashev and Werb 1998; Schuster and Krieglstein 2002). Finally, the Wnt signaling pathway is generally involved in cell development processes, such as cell-fate determination and cell differentiation (Logan and Nusse 2004). It also plays a role in immature T-cell and B-cell proliferation, migration of peripheral T-cells, and modulation of antigen presenting cells such as dendritic cells (Staal, Luis et al. 2008).

3.2.5 Concluding remarks

In this study, we applied an X-tailored analysis pipeline to 16 different GWAS datasets (Table 3.1), discovered and replicated several genes associated with autoimmune disease risk. Multiple additional lines of evidence point to some of these genes having immune-related functions,

including expression in immune-related tissues (Figs. 3.2-3.3), in addition to enrichment for these genes and their interacting partners in immune related pathways (Table 3.3). Beyond immune function, several of the genes we associated with disease risk (*IL13RA1, ARHGEF6, MCF2*) are also involved in regulation of apoptosis. Apoptosis has long been suspected of playing a role in AID (Eguchi 2001; Kawakami and Eguchi 2002; Mason, Lin et al. 2013) and shows strong evidence for involvement in the etiology of vitiligo[90], psoriasis (Weatherhead, Farr et al. 2011) and rheumatoid arthritis (Li, Ma et al. 2014). Our analyses also highlight the sex specific nature of associated disease risk genes shedding light on the sexual dimorphism of some autoimmune and immune-mediated diseases.

The X chromosome has received special attention in GWAS during the past year (Conde, Foo et al. 2013; Wise, Gyi et al. 2013; Konig, Loley et al. 2014; Tukiainen, Pirinen et al. 2014). Our results highlight chromosome X's contribution to sex-differences in disease risk and yield new avenues for potential functional follow-up. More generally, our study illustrates that with the right tools and methodology, new discoveries regarding the role of X in complex disease and sexual dimorphism can be made, even with existing, array-based GWAS datasets. To enable researchers to make many additional such discoveries by analyzing this unique chromosome in the context of existing and emerging genome-wide association studies, we have released our software for handling chromosome X (Chang, Gao et al.), which we provide as an extension of PLINK (Purcell, Neale et al. 2007). Further expansions of this initial software can take unique X-related features to further develop X-tailored methods such as methods that rely on X-inactivation and on the availability of phased X haplotypes in males.

3.3 Materials and Methods

3.3.1 Datasets

We obtained the following datasets from dbGaP: ALS Finland (Laaksovirta, Peuralinna et al. 2010) (phs000344), ALS Irish (Cronin, Berger et al. 2008) (phs000127), CIDR Celiac disease (Ahn, Ding et al. 2012) (phs000274), MScc (Baranzini, Wang et al. 2009) (phs000171), Vitiligo GWAS1 (Jin, Birlea et al. 2010) (phs000224), NIDDK CD (Duerr, Taylor et al. 2006) (phs000130), CASP (Nair, Duffin et al. 2009) (phs000019), and GENEVA T2D (Qi, Cornelis et al. 2010) (phs000091). The Vitiligo GWAS2 (Jin, Birlea et al. 2012) dataset was provided by R.S. Both vitiligo datasets contained case data only. Therefore, we obtained the following additional control datasets from dbGaP: PanScan (Amundadottir, Kraft et al. 2009; Petersen, Amundadottir et al. 2010) (phs000206), National Institute on Aging Alzheimer's study (Lee, Cheng et al. 2008) (phs000168), CIDR bone fragility (Estrada, Styrkarsdottir et al. 2012) (phs000138), COGA (Bierut, Saccone et al. 2002) (phs000125), and SAGE (Bierut, Saccone et al. 2002; Bierut 2007; Bierut, Strickland et al. 2008) (phs000092). Only samples with the "general research consent" designation in the control datasets were used as controls for studying vitiligo. These samples were randomly distributed between the two vitiligo datasets.

Additional datasets were obtained from the Wellcome Trust Case Control Consortium (WT): all WT1 (The Wellcome Trust Case Control Consortium 2007) datasets, WT2 ankyolosing spondylitis (AS) (Evans, Spencer et al. 2011), WT2 ulcerative colitis (UC) (Barrett, Lee et al. 2009) and WT2 multiple sclerosis (MS) (Sawcer, Hellenthal et al. 2011). In order to run meta-analysis and independently replicate signals, we ensured that none of these datasets had overlapping controls. To accomplish this, we recruited additional control data from the WT1

hypertension (HT), bipolar (BP), and cardiovascular disease (CAD) case data. These samples were randomly distributed to the four WT1 datasets, though only BP samples were used as controls for WT1 T2D due to potential shared disease etiology between T2D, CAD and HT. The WT1 National Birth Registry (NBS) control data was also randomly distributed to the four WT1 datasets. We randomly distributed the 58 Birth Cohort (58BC) control samples, along with any new NBS samples not present in the WT1 data, between WT2 datasets.

3.3.2 Quality Control (QC)

Our pipeline for X-wide association studies (XWAS) begins with a number of quality control steps, some of which are specific to the X chromosome. First, we removed samples that we inferred to be related, had > 10% missing genotypes, and those with reported sex that did not match the heterozygosity rates observed on chromosome X. We additionally filtered variants with >10% missingness, variants with a minor allele frequency (MAF) < 0.005, variants for which missingness was significantly correlated with phenotype ($P<1x10^{-4}$). X-specific QC steps included filtering variants not in Hardy-Weinberg equilibrium in females ($P<1x10^{-4}$), removing variants that had significantly different MAF between males and females in control individuals; and removal of the pseudoautosomal regions (PARs).

3.3.3 Correction for population stratification

To assess and adjust for potential population stratification we ran principal component analysis (PCA) using EIGENSOFT (Patterson, Price et al. 2006) after pruning for linkage disequilibrium (LD) and removing large LD blocks (Novembre, Johnson et al. 2008). Individuals inferred to be of non-European ancestry were removed from all subsequent analysis. For the datasets analyzed

here (of European ancestry), we found that correction for population stratification is more accurate when based on the autosomes than on X alone due to the smaller number SNPs used to infer structure based on X. This observation holds as long as enough autosomal principal components (PCs) are considered. Thus, in our subsequent analyses, only the first ten autosomal derived PCs were used to assess and correct for population stratification. Sex-biased demographic events though, including sex-differential population structure of males and females, such as events proposed for human populations (Hammer, Mendez et al. 2008; Keinan and Reich 2010; Heyer, Chaix et al. 2012), are expected to lead to differential population structure on X and the autosomes. Hence, the problem of population stratification can be different between the two genomic compartments. In theory, this suggests that correction for population stratification in XWAS should be based on inference of population structure utilizing the X chromosome alone. Given this, we anticipate cases—in other populations or where more data is available for X—in which correction for population stratification based on X alone could potentially be more accurate for XWAS.

3.3.4 Imputation

Imputation was carried out with IMPUTE2 (Howie, Donnelly et al. 2009) version 2.2.2 based on 1000 Genomes Project (Abecasis, Auton et al. 2012) whole-genome and whole-exome (October 2011 release) haplotype data. One of the features added in the second version of impute (IMPUTE2) is the assumption of a 25% reduction in the effective population size (*Ne*) when imputing variants on the X chromosome. As recommended by the authors IMPUTE2, *Ne* was set to 20,000 and variants with MAF in Europeans < 0.005 were not imputed. Based on the output of IMPUTE2, we excluded variants with an imputation quality < 0.5 and variants that did not pass

the above QC criteria (see *Quality Control*). Table 3.1 displays the number of SNPs we considered in each dataset following imputation and these additional QC steps.

3.3.5 Single marker association analysis

In the first test we assume complete X-inactivation and similar effect size between males and females. While females are considered to have 0, 1, or 2 copies of an allele (as in the autosomes), males are considered to have 0 or 2 copies of the same allele. Thus, male hemizygotes are equivalent to female homozygous states. This test is currently implemented in PLINK (Purcell, Neale et al. 2007) as the *–xchr-model 2* option, termed FM_{02} in this study. In the second test, male and female data are analyzed separately (with males coded as either having 0 or 2 copies of an allele as above). The female only and male only measures of significance are then combined using either Fisher's (Fisher 1925) method or a weighted Stouffer's method (Stouffer, Suchman et al. 1949), with weighting determined by sample size (Willer, Li et al. 2010) to obtain the $FM_{F.comb}$ and the $FM_{S.comb}$ test association p-values. Fisher's method combines the p-values themselves, while Stouffer's method combines test statistics, taking into account both the sample size and direction of effect for males and females. Ten PCs were added as covariates to account for potential population stratification. Principal component covariates were not added to the regression model for the amyotrophic lateral sclerosis (ALS) Finland, ALS Irish, and CASP datasets as no inflated p-values were observed in these studies (Supplementary Fig. 3.S1).

3.3.6 Gene-based analysis

Gene-based association analysis was carried out in the general framework of VEGAS (Liu, McRae et al. 2010). We briefly summarize the method here, though a more detailed description

can be found in (Liu, McRae et al. 2010). As SNPs in a gene are in closer proximity with each other, they are likely to be in LD and thus have correlated test statistics. VEGAS accounts for this correlation by utilizing the LD between SNPs in a gene to derive the distribution of test statistics (Liu, McRae et al. 2010). More specifically, *n* statistics are then randomly drawn from a multivariate normal distribution with a mean of 0 and a *n* x *n* covariance matrix corresponding to the pairwise LD between SNPs mapped to the gene, where *n* represents the number of SNPs in a gene. These *n* statistics are then combined via summation. Here, we have implemented a slight modification to this procedure. In this study, we combined p-values derived from the simulated test statistics with either the truncated tail strength (Jiang, Zhang et al. 2011) or the truncated product (Zaykin, Zhivotovsky et al. 2002) method, which have been suggested to be more powerful than other tests in some scenarios (Huang, Chanda et al. 2011; Ma, Clark et al. 2013). The gene-based p-value was calculated as the proportion of simulated statistics that were as extreme or more extreme than the observed statistic. To increase time efficiency of the simulation procedure, adaptive simulations were implemented as in VEGAS (Liu, McRae et al. 2010). A list of X-linked genes and their positions was obtained from UCSC "knownCanonical" transcript ID track (http://genome.ucsc.edu/cgi-bin/hgTrackUi?db=hg19&g=knownGene). SNPs were mapped to a gene if they were within 15 kilobases (kb) of a gene's start or end positions.

3.3.7 Sex-difference analysis

The difference in the effect size between males and females at each SNP was assayed using a t-statistic as calculated below (Randall, Winkler et al. 2013):

$$t = \frac{\log(OR_{male}) - \log(OR_{female})}{\sqrt{SE_{male}^2 + SE_{female}^2 - 2rSE_{male}SE_{female}}}$$

SE is in the standard error in males or females, and *r* is the Spearman rank correlation coefficient between log(OR_{male}) and log(OR_{female}) across all X-linked SNPs. The odds ratio in males is estimated with 02 coding for male genotypes as in the FM_{02} test. This test is most powered to detect variants with opposing effects in males versus females, though it will also capture cases where the effects are in the same direction though significantly different in magnitude.

3.3.8 Gene expression analysis

Whole blood gene expression data for 881 samples (409 males, 472 females) from the Rotterdam Study III (Hofman, Breteler et al. 2009) was downloaded from Gene Expression Omnibus (Barrett, Wilhite et al. 2013) (accession GSE33828). Expression data was available for 802 of the genes studied in our XWAS. For each gene, we tested for differential expression between males and females using the Wilcoxon rank sum test across individuals and applied Bonferroni correction to its p-values. Using a hypergeometric test, we also assayed whether the 802 X-linked genes analyzed in our study are more often differentially expressed between males and females as compared to all genes genome-wide. In addition, we assessed how many of the genes that were associated and replicated (20 genes with expression data) showed significant differential expression between males and females (after Bonferroni correction for the number of associated and replicated genes). To assess co-expression between X-linked genes, we calculated the non-parametric Spearman correlation coefficient between the expression of each pair of genes across the set of 881 individuals. Enrichment of significant co-expression within the set of 20 associated genes as compared to all 802 genes was tested using a hypergeometric test.

For analysis of tissue-specific gene expression, we obtained the Human GNF1H tissue-specific

expression dataset (Su, Wiltshire et al. 2004) via the BioGPS website (Wu, Macleod et al. 2013). After excluding fetal and cancer tissues, we were left with expression data across 74 tissues for 504 of the genes studied in our XWAS, including 14 of the genes with evidence of association (Fig. 3.1). For each gene, we obtained a normalized z-score value for its expression in each tissue by normalizing its expression by the average and standard deviation of the expression of that gene across all tissues.

3.3.9 Network analysis

A network of interacting genes was assembled in GeneMANIA using confirmed and predicted genetic and protein interactions (Warde-Farley, Donaldson et al. 2010) with a seed list of the 22 protein-coding genes within the list of associated genes (Fig. 3.1). Up to 100 genes were added with a maximum of 20 attributes. Scores for interactions were weighted equally by network. This scoring allows for querying interactions between genes while minimizing bias from obtaining more hits in well-studied pathways. A list of unique genes within this interactome was extracted as input to WebGestalt (Zhang, Kirov et al. 2005; Wang, Duncan et al. 2013) to discover the ten most significantly enriched pathways in the KEGG (Kanehisa and Goto 2000) database. Enrichment was assessed with the hypergeometric test (Wang, Duncan et al. 2013) and reported p-values were adjusted for multiple testing using the Benjamini-Hochberg correction. Pathways were required to have a minimum of two genes.

3.3.10 Gene-set analysis

We additionally tested whether SNPs in a set of genes were collectively associated with disease. To accomplish this, we modified the gene-based analysis above to draw from multiple

multivariate normal (MVN) distributions, each with their own covariance matrix corresponding to the LD between SNPs in each gene within the gene-set. Comparing p-values derived from 100 phenotypic permutations to this simulation procedure revealed highly correlated significance values (Supplementary Fig. 3.S4-3.S5). We thus only present results from our simulation procedure.

We manually curated a set of immune-related genes from the KEGG (Kanehisa and Goto 2000) pathways and Gene Ontology (GO) (Ashburner, Ball et al. 2000) Biological Function categories. To do this we mined the KEGG and GO databases using 15 and 14 categories, respectively, that are particularly relevant for autoimmune response. We subsequently removed eight genes from this list that we felt were either too generalized (e.g. cell cycle genes) or too specific (e.g. F8 and F9 blood coagulation genes) to obtain a final list of 27 genes (Supplementary Table 3.S9). The Panther immune gene set was obtained by including genes in the category of "immune system processes" in the Panther database (Thomas, Campbell et al. 2003). The XY homolog gene set was obtained from Wilson-Sayres and Makova (Wilson Sayres and Makova 2013).

3.4 Supplementary Text

The single nucleotide polymorphism (SNP) association analysis, including imputed SNPs, identified 42 SNPs significantly associated with their respective disease following a conservative Bonferroni correction for the number of tests (Supplementary Fig. 3.S2a, Table 3.S1). Of these, 14 SNPs in the same locus form a clear peak (Supplementary Fig. 3.S2b) in their association with vitiligo (Vitiligo GWAS1 dataset). Vitiligo is a common autoimmune disorder in which the destruction of melanocytes (pigment producing cells located in the basal epidermis) results in depigmented skin. The associated locus is 17 kilobases (kb) away from a weakly expressed retrotransposed gene (retro-*HSPA8*) that is of 98% similarity to its parent gene, *HSPA8*, on chromosome 11. *HSPA8* encodes a member of the heat shock protein 70 family and functions as a chaperone to bind nascent polypeptides and enable correct folding. Heat shock proteins have been previously implicated in autoimmune disease(Winfield and Jarjour 1991; Rauch, San Martin et al. 1995; Ludwig, Stahl et al. 1999; Naumann, Hempel et al. 2001; Routsias and Tzioufas 2006). In particular, a role for inducible heat shock protein 70 has been suggested in vitiligo(Mosenson, Zloza et al. 2012; Abdou, Maraee et al. 2013; Mosenson, Eby et al. 2013). Though this region did not replicate in our second vitiligo dataset, the biological relevance of this region warrants further investigation in a larger, better powered replication study. Another clear association peak was observed for the Wellcome Trust Case Control Consortium 2 ulcerative colitis (WT2 UC) (Supplementary Fig. 3.S1c) for intronic variants of *BCOR*. *BCOR* encodes a co-repressor of *BCL-6*, which regulates apoptosis(Huynh, Fischle et al. 2000). However, none of these candidate associations replicated in other GWAS datasets for the same or related diseases,

possibly due to small sample sizes and thus insufficient power (Table 3.1).

There is abundant evidence that many autoimmune and immune-related disorders share some genetic etiology(Sirota, Schaub et al. 2009; Cotsapas, Voight et al. 2011; Solovieff, Cotsapas et al. 2013; Chang and Keinan 2014). While these studies have focused on autosomal variants, this may also be the case for chromosome X. We therefore used PLINK(Purcell, Neale et al. 2007) to perform a fixed-effects meta-analysis on several subsets of our diseases. We performed these analyses using the p-values generated from the FM_{02} test. Specifically, we applied this analysis to the following disease sets (see Table 3.1 for dataset designations): (i) all classic autoimmune diseases: CIDR celiac disease, WT1 CD, WT1 RA, WT1 T1D, WT2 UC, WT2AS, WT2 MS, CASP; (ii) WT2 AS, WT1 RA, WT1 T1D, CIDR celiac disease(Sirota, Schaub et al. 2009); (iii) classical neurological disorders: ALS Finland, ALS Irish, WT2 MS, MS case control (iv) diabetes: WT1 T1D, WT1 T2D; (v) irritable bowel disease: WT1 CD, NIDDK CD, WT2 UC; (vi) and skin related disorders: Vitiligo GWAS1 and CASP.

We found 4 regions containing SNPs with $P < 1x10^{-4}$ in three of the disease groups (Supplementary Table 3.10a-c). While not significant after a conservative Bonferroni correction for the number of SNPs tested on X, the most significant SNP in the inflammatory bowel disorder disease group ($P= 1.73x10^{-5}$) is located ~38 kb from *CD40LG*, which encodes a protein expressed on the surface of T-cells (Supplementary Table 3.10a-c). Furthermore, one of the most significant SNPs in the psoriasis and vitiligo meta-analysis (rs11797576, $P = 7.13x10^{-5}$) is located 50 kb from *EGFL6*, which encodes an epidermal growth factor (Supplementary Table 3.10a-c).

We further tested for a significant difference in effect size between males and females (see Methods in the main text). We found one significant SNP (rs200718, $P = 1.51 \times 10^{-7}$) in Vitiligo GWAS1. Association of this SNP was not replicated in Vitiligo GWAS2 due to its very low minor allele frequency (MAF < 0.003). We thus assayed whether the nearby SNP rs5976539, in moderate LD with rs200718 (D' = 0.306), was associated in Vitiligo GWAS2, but did not find evidence for a significant difference in effect size between males and females (P = 0.920).

3.5 Figures

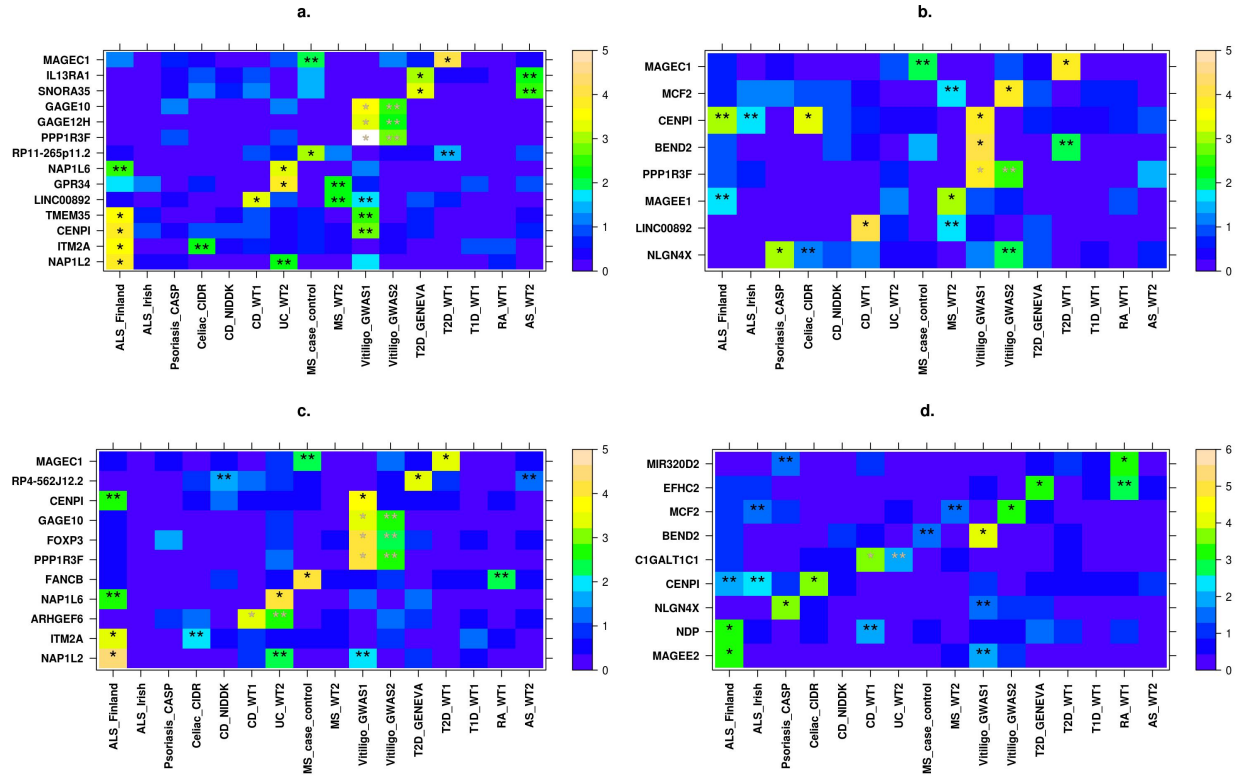

Figure 3.1. X-linked genes associated with autoimmune disease risk.
All genes that showed evidence of association in a gene-based test ($P<10^{-3}$), and replicated in another dataset are shown for the a) $FM_{S.comb}$ b) $FM_{F.comb}$ c) FM_{02} and d) sex-differentiated effect size tests (Materials and Methods). Dataset names, as described in Table 3.1, are displayed on the *x-axis* and gene names on the *y-axis*. For each gene, the more significant p-value of the truncated tail strength and truncated product methods is displayed on a –log10 scale according to the enclosed color scale. A "*" represents the discovery dataset, while "**" indicates the replication dataset/s. These appear in grey when the discovery and replication are in datasets of the same disease (or across the related Crohn's disease and ulcerative colitis). Numerical values corresponding to this table are presented in Table 3.S4 and 3.S6.

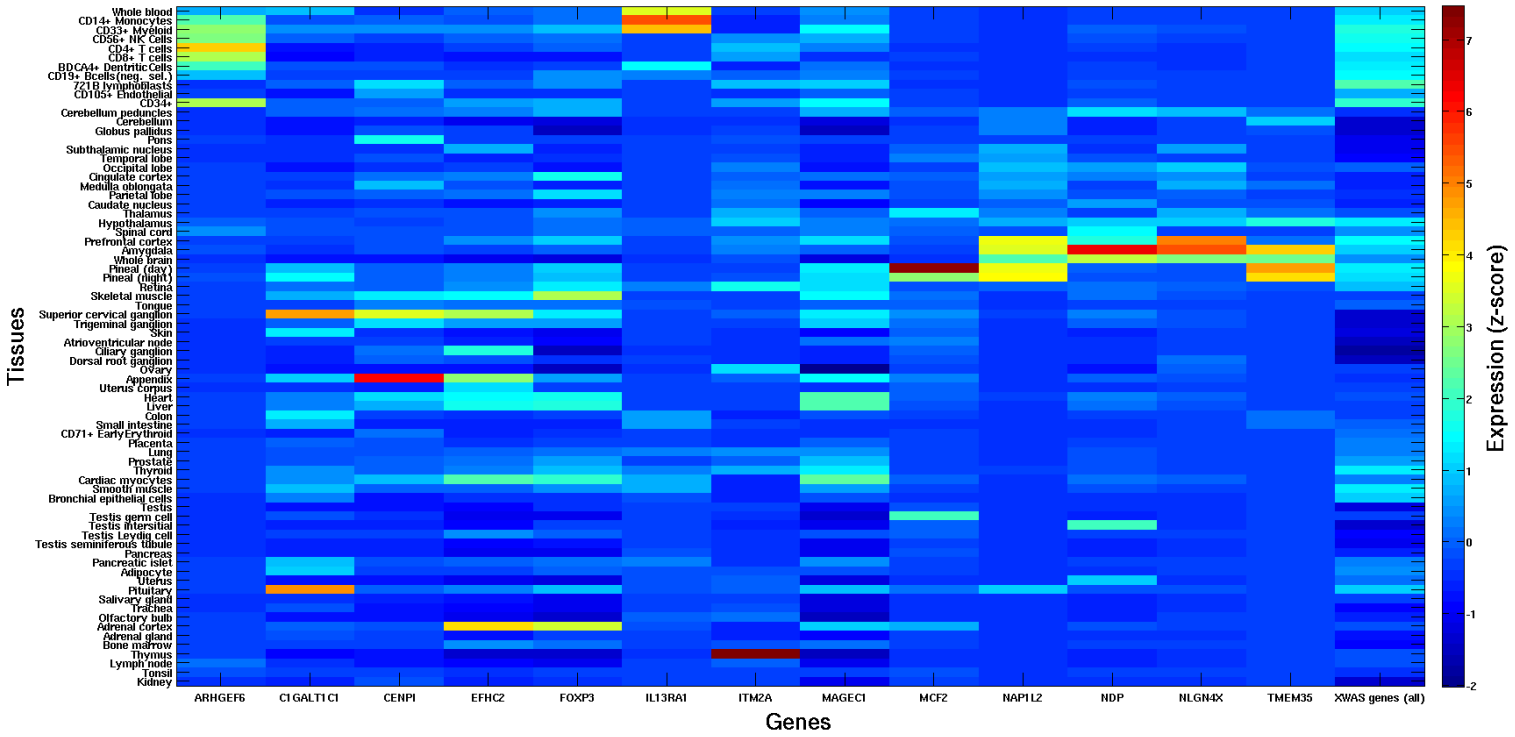

Figure 3.2. X-linked disease risk genes are differentially expressed between tissues.
X-axis presents 13 out of the associated X-linked genes for which gene expression data was available for analysis. For each, a z-score is presented for the deviation of expression in each of 74 tissues (y-axis) from the average expression of that gene across all tissues (Materials and Methods). For comparison, the last column shows average expression in each tissue across all X-linked genes that were tested as part of our analyses. Several associated genes exhibit significantly higher expression in immune-related tissues (see main text).

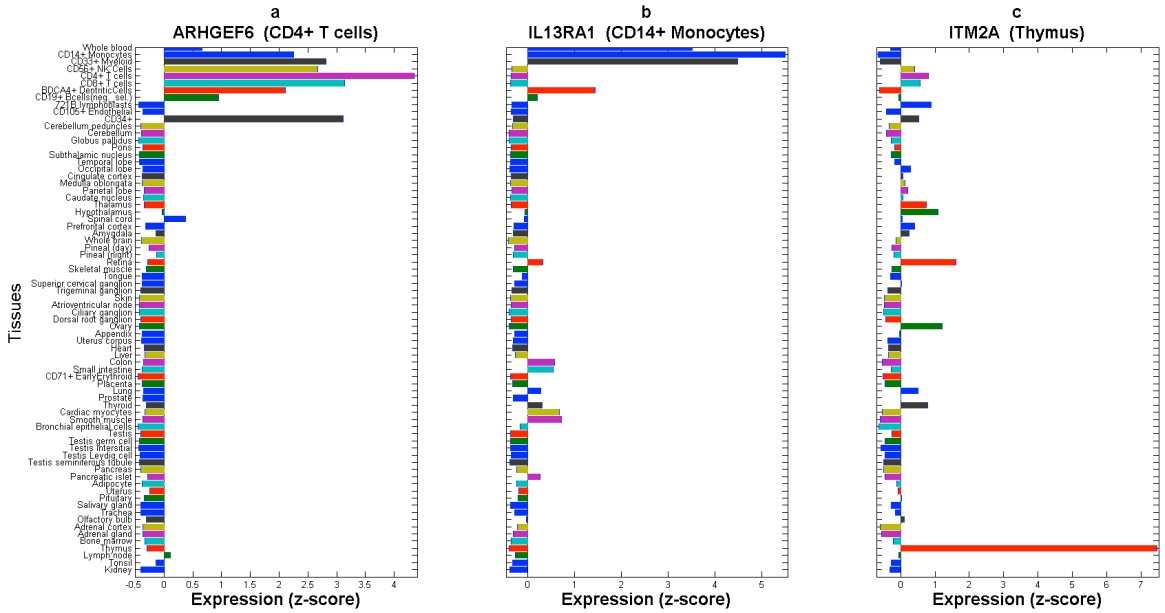

Figure 3.3. Three X-linked disease risk genes show high expression in immune-related tissues and cells.

ARHGEF6, *IL13RA1*, and *ITM2A* show expression greater than 4 standard deviations above the average expression of these genes in T-cells (highest in CD4+), CD14+ monocytes, and the thymus, respectively (Fig. 3.2). Y-axis follows the respective tissues from Figure 3.2 and x-axis denotes a z-score for the deviation of expression in each tissue from the average expression of that gene. The title of each panel includes the name of the gene and the tissue with the highest expression for that gene.

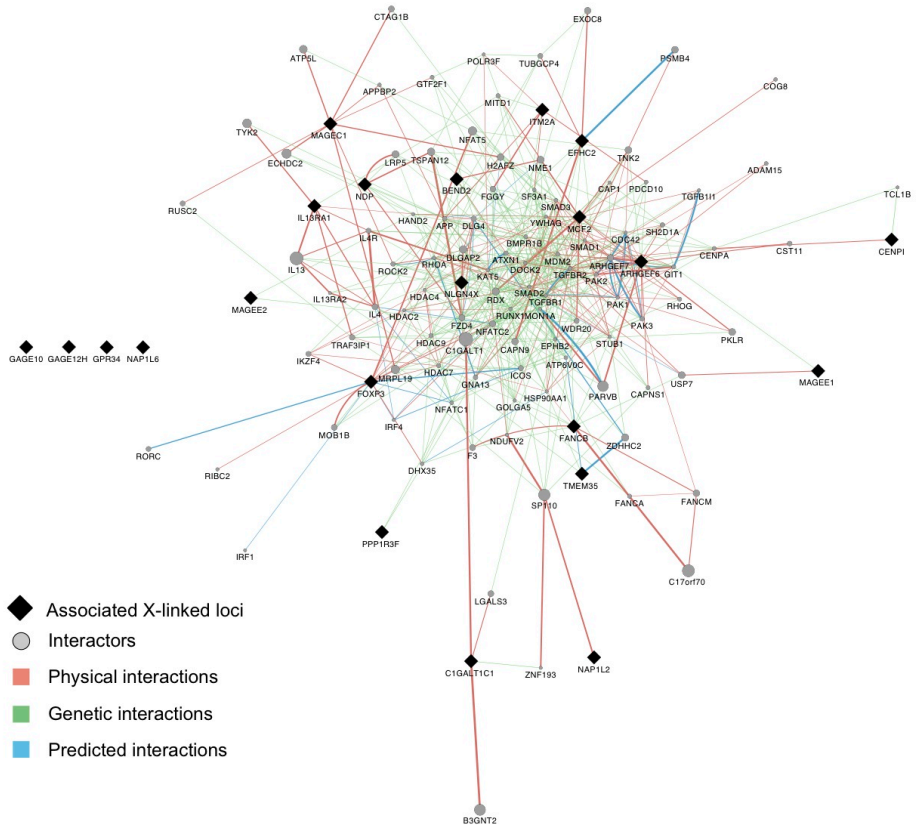

Figure 3.4. Interactome of X-linked disease risk genes.
All 22 associated X-linked protein-coding genes (Fig. 3.1), denoted by black diamonds, together with genes that interact with them. *Physical interactions* refer to documented protein-protein interactions. *Genetic interactions* represent genes where perturbations to one gene affect another. *Predicted interactions* were obtained from orthology to interactions present in other organisms (Warde-Farley, Donaldson et al. 2010). All but three genes associated with AID and DPACs share interacting partners according to known and predicted interactions (Materials and Methods).

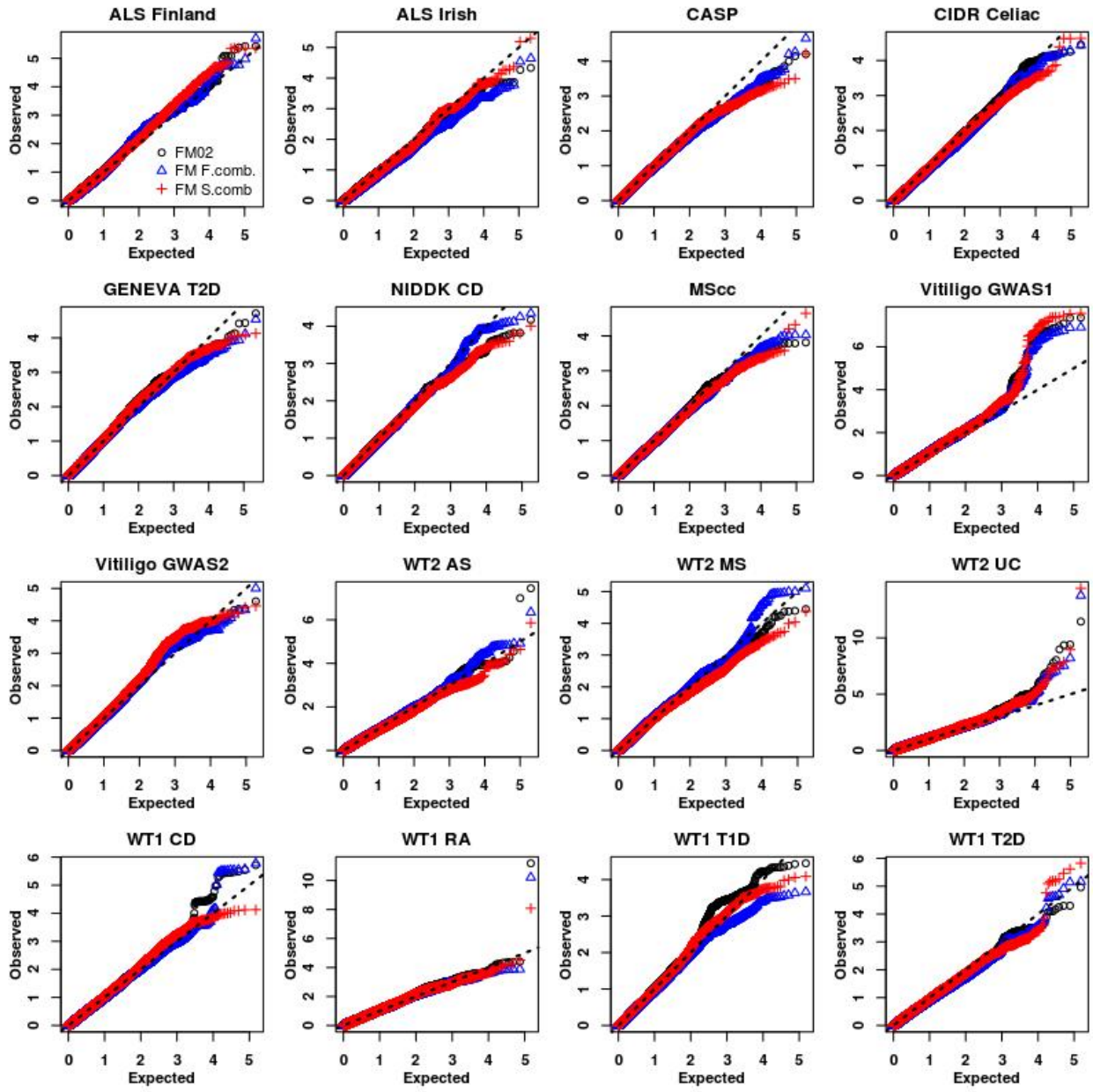

Figure 3.S1. QQ-plots for single marker association tests.
Blue triangles denote association p-values for the $FM_{F.comb}$ test, red crosses denote p-values for the $FM_{S.comb}$, while the black points denote association p-values for the FM_{02} test. P-values are plotted in log scale.

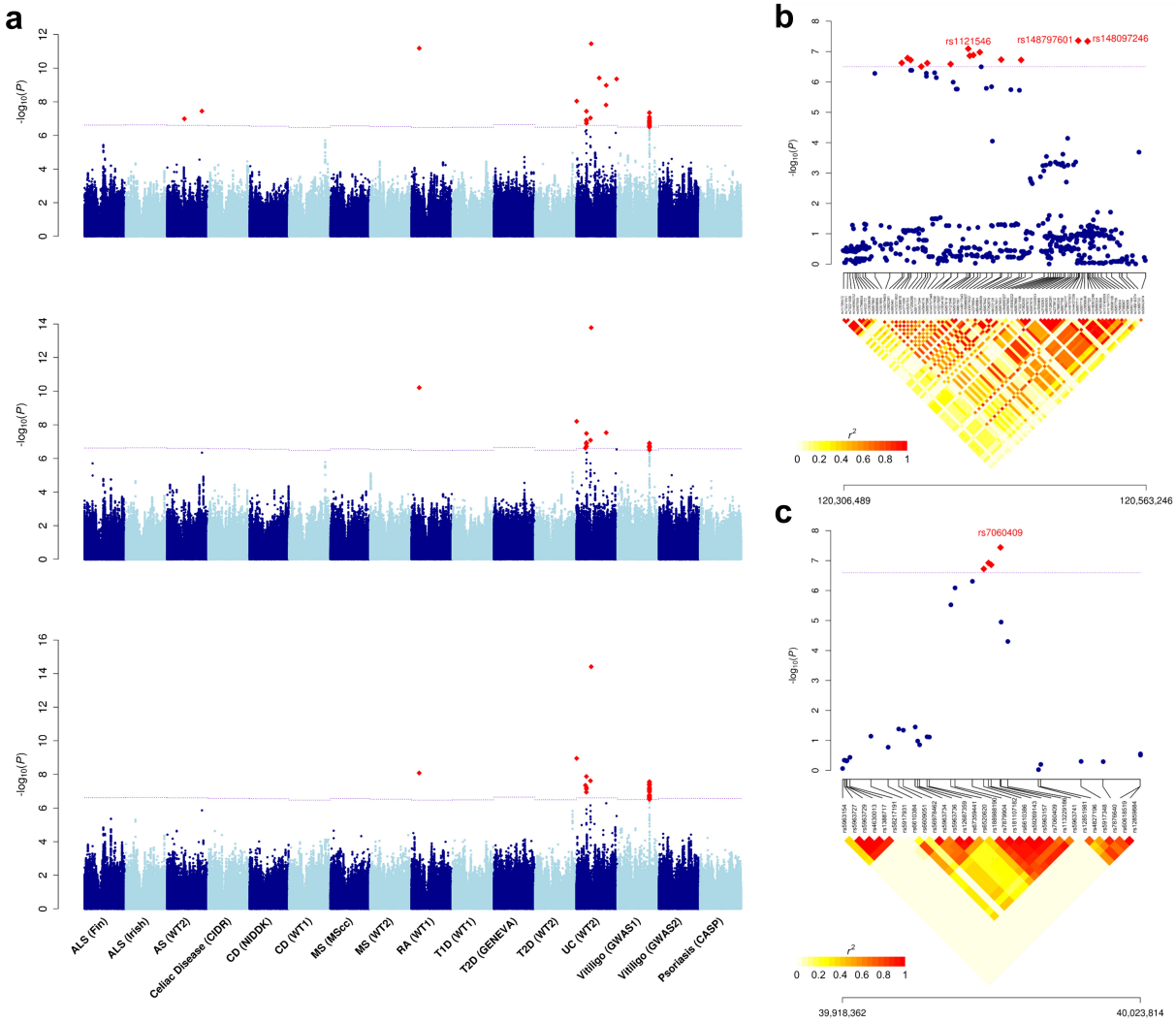

Figure 3.S2. Significant SNP associations.
(a) A Manhattan plot of the nominal p-values for the FM_{02} (upper), $FM_{F.comb}$ (middle), and $FM_{S.comb}$ (lower) tests of association for chromosome X SNPs in the 16 datasets. The dotted purple lines correspond to the significance threshold for each dataset. The significant associations are shown as red diamonds. (b-c) Regional association plots of the association results and LD for (b) Vitiligo GWAS1 data set and (c) WT2 UC data set. Upper: the purple dotted line corresponds to the significance threshold, and the significant results are shown as red diamonds. Lower: LD structure was plotted using a revised version of the *snp.plotter* software (Luna and Nicodemus 2007). Due to the large number of SNPs in the associated region of Vitiligo GWAS1, only every 1 in 10 of the non-significantly associated SNPs are shown.

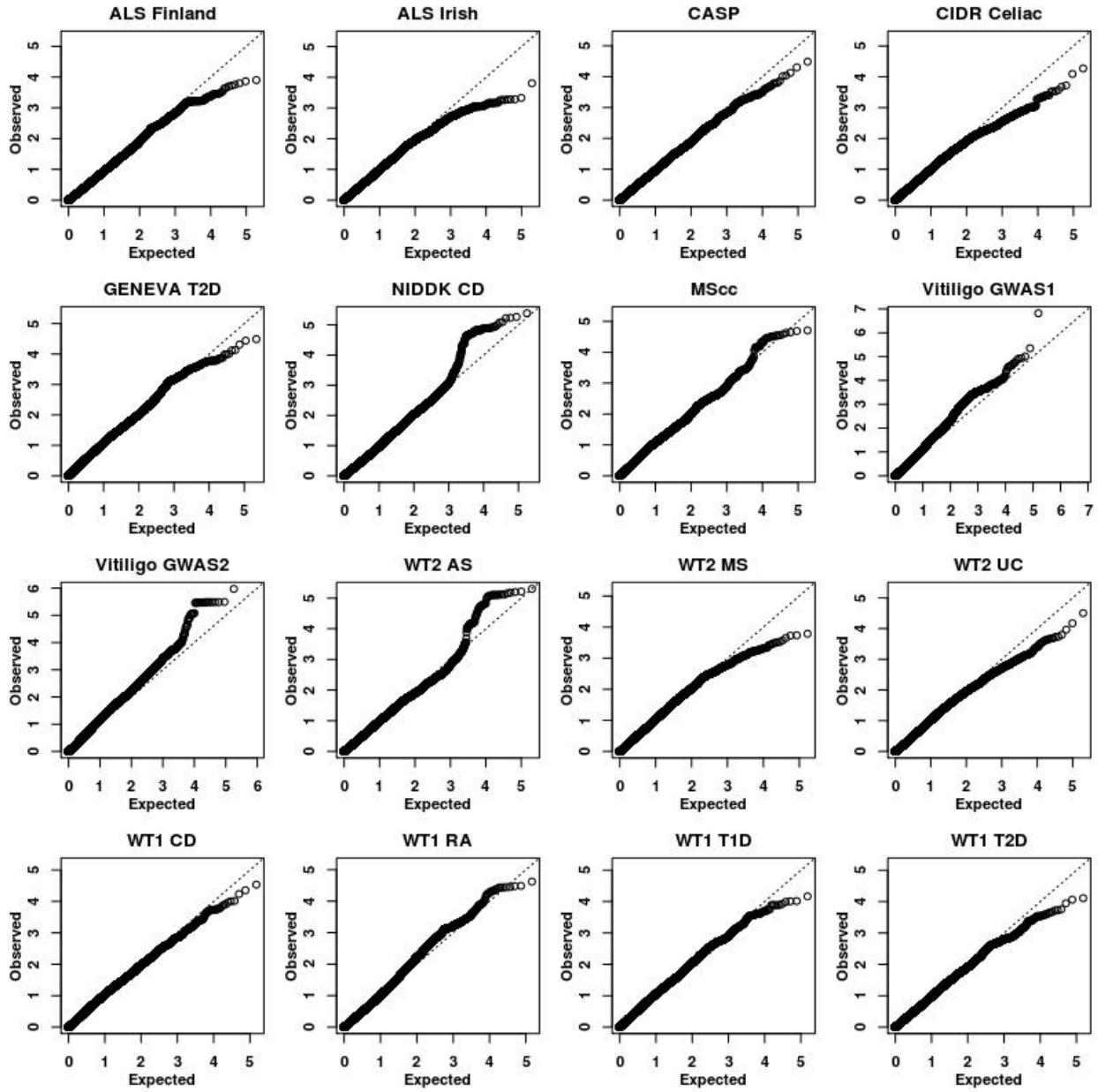

Figure 3.S3. QQ-plots for test of sex-differentiated effect size.
Similar to Figure 3.S1, where p-values are now displayed for the test of differential effect size between males and females.

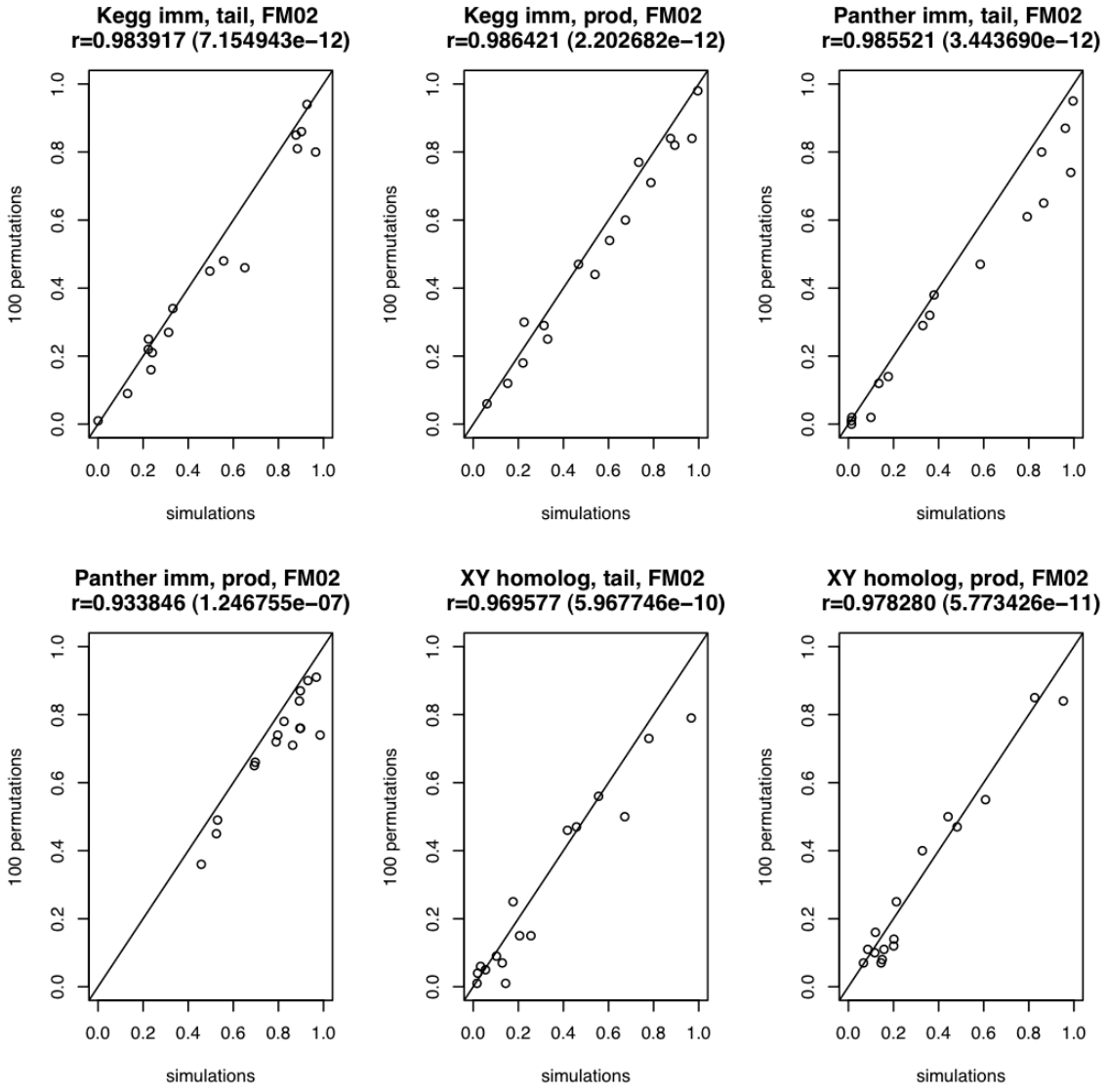

Figure 3.S4. Comparison between simulation derived and permutation derived p-values for the gene-set association analysis using the FM_{02} test statistic.
r represents Pearson's correlation coefficient and the significance of the correlation is indicated in the parentheses in scientific notation.

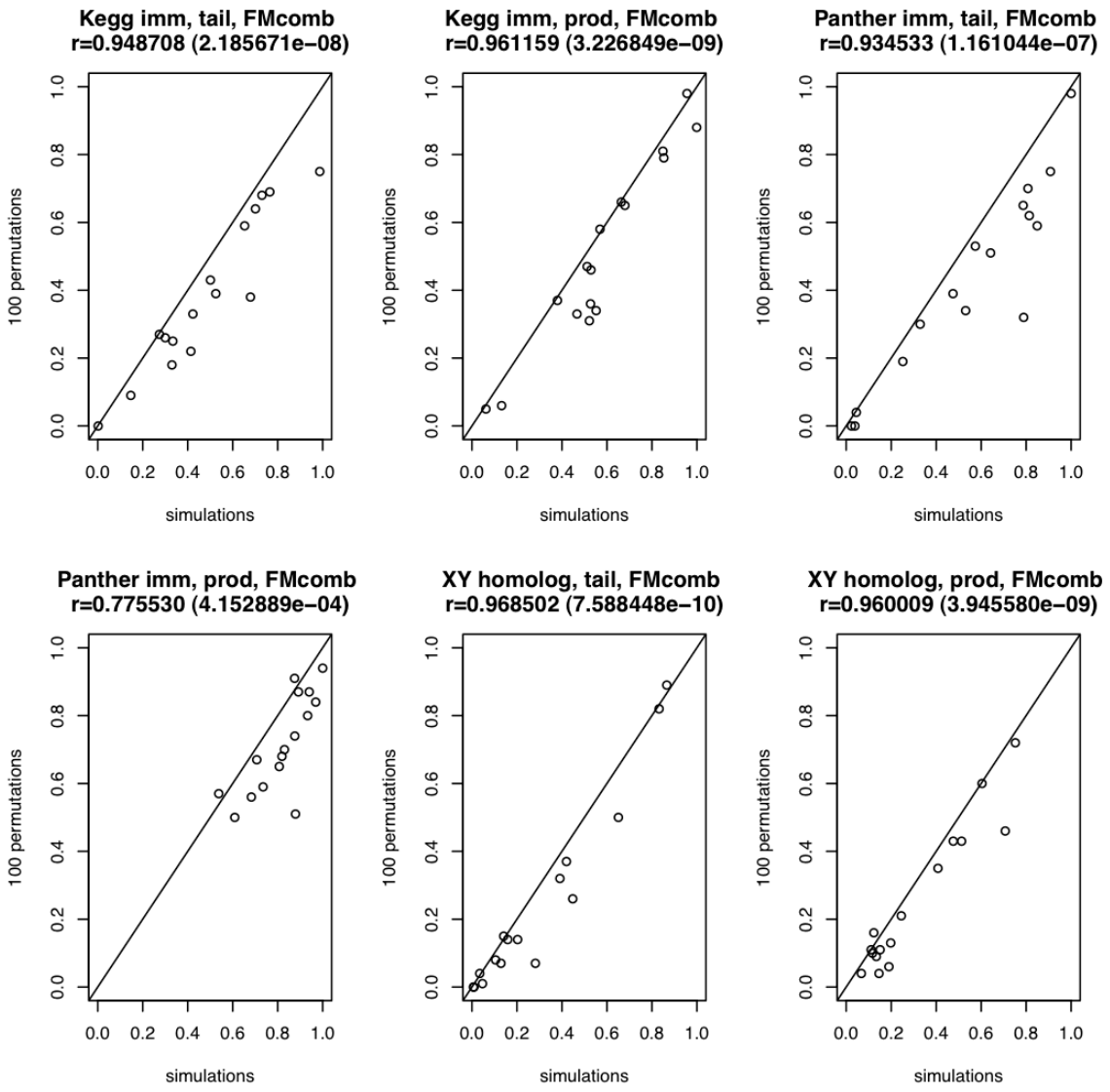

Figure 3.S5. Comparison between simulation derived and permutation derived p-values for the gene-set association analysis using the $FM_{F.comb}$ test statistic.
Similar to Figure 3.S5 where test statistics for the $FM_{F.comb}$ test are now displayed.

3.6 Tables

Dataset	Disease	# SNPs	# Genes	# Cases	# Controls
ALS Finland	Amyotrophic Lateral Sclerosis (ALS)	207,947	970	400	490
ALS Irish	Amyotrophic Lateral Sclerosis (ALS)	219,300	967	221	210
Psoriasis CASP	Psoriasis	184,246	953	1,209	1,271
Celiac Disease CIDR	Celiac Disease	187,284	962	1,576	504
CD NIDDK	Crohn's Disease (CD)	176,072	837	791	922
CD WT1*	Crohn's Disease (CD)	150,275	930	1,592	1,701
UC WT2*	Ulcerative Colitis (UC)	196,781	963	2,341	1,699
MS case control	Multiple Sclerosis (MS)	183,954	842	943	851
MS WT2*	Multiple Sclerosis (MS)	169,707	962	2,666	1389
Vitiligo GWAS1	Vitiligo	157,676	958	1,391	4,521
Vitiligo GWAS2	Vitiligo	187,688	962	415	2,552
T2D GENEVA	Type-2 Diabetes (T2D)	220,752	971	2,515	2,850
T2D WT1*	Type-2 Diabetes (T2D)	152,996	927	1,811	1,668
T1D WT1*	Type-1 Diabetes (T1D)	152,304	926	1,867	1,714
RA WT1*	Rheumatoid Arthritis (RA)	146,907	925	1,772	1,709
AS WT2*	Ankyolosing Spondylitis (AS)	200,042	966	1,472	1,260

Table 3.1. GWAS datasets.
For each of the case-control datasets analyzed in this study, the table lists its name, the disease considered, the number of X-linked SNPs (*# SNPs*), which include imputed SNPs, and the number of genes tested in the gene-based test (*# Genes*). The number of individuals *(# Cases* and *# Controls*) represents the number of samples following QC. All datasets consist of individuals of European ancestry. Though ALS and T2D are not conventionally considered as autoimmune diseases, we have included datasets of these diseases due to recent studies pointing to an autoimmune component to their etiology (Pagani, Gonzalez et al. 2011; Itariu and Stulnig 2014). *As control individuals overlap across these datasets, we only considered non-overlapping subsets of them for each of the diseases studied here (Materials and Methods). The size of these subsets is indicated under *# Controls*.

Dataset	Statistic	P-value
XY homologs gene set		
Psoriasis CASP	$FM_{F.comb}$	**0.0088**
Celiac disease CIDR	$FM_{F.comb}$	0.0467
Vitiligo GWAS1	$FM_{F.comb}$	**0.0063**
Vitiligo GWAS1	FM_{02}	0.0329
Vitiligo GWAS2	$FM_{F.comb}$	0.0346
CD NIDDK	FM_{02}	0.017
CD WT1	FM_{02}	0.0234
T1D WT1	$FM_{S.comb}$	0.0302
Panther immune gene set		
Vitiligo GWAS1	FM_{02}	**0.0154**
Vitiligo GWAS1	$FM_{F.comb}$	0.0387
Vitiligo GWAS1	$FM_{S.comb}$	**0.0081**
Vitiligo GWAS2	FM_{02}	**0.0142**
Vitiligo GWAS2	$FM_{F.comb}$	0.0448
Vitiligo GWAS2	$FM_{S.comb}$	**0.0127**
T2D GENEVA	$FM_{S.comb}$	**0.0073**
KEGG/GO immune gene set		
Vitiligo GWAS1	$FM_{F.comb}$	**0.002**
Vitiligo GWAS1	$FM_{s.comb}$	**1.64×10^{-4}**

Table 3.2. Gene-set associations.
Three curated gene sets were tested for association to diseases. Datasets with p-values < 0.05 are displayed, with bold p-values indicating significant association after multiple testing correction. The minimum of the truncated tail strength method and the truncated product method are displayed.

Pathway	Genes	P-value
Regulation of actin cytoskeleton	*PAK1, RHOA, PAK3, CDC42, ARHGEF6, SOS1, ARHGEF7, PAK2, RDX, GIT1, GNA13, TIAM1, ROCK2, FGD1*	5.55×10^{-14}
T-cell receptor signaling pathway	*PAK1, RHOA, PAK3, CDC42, SOS1, PAK2, IL4, NFATC2, NFATC1, ICOS, NFAT5*	2.75×10^{-13}
Axon guidance	*PAK1, RHOA, PAK3, EPHB2, CDC42, NFATC2, NFATC1, NFAT5, ROCK2*	4.97×10^{-11}
Wnt signaling	*SMAD3, SMAD2, RHOA, FZD4, LRP5, NFATC2, NFATC1, NFAT5, ROCK2*	4.74×10^{-9}
Systemic lupus erythematosus	*H2AFZ, H2AFJ, HIST1H2AH, HIST2H2AB, HIST1H2AJ, HIST3H2A, HIST1H2AD*	4.34×10^{-8}
Chemokine signaling	*PAK1, RHOA, CDC42, SOS1, GNB1, TIAM1, DOCK2, ROCK2*	4.52×10^{-7}
Focal adhesion	*PAK1, PARVB, RHOA, PAK3, CDC42, SOS1, PAK2, ROCK2*	6.28×10^{-7}
TGF-beta signaling	*SMAD3, SMAD2, RHOA, TGFBR2, ROCK2, BMPR1B*	7.87×10^{-7}
Pathways in cancer	*SMAD3, SMAD2, RHOA, MDM2, CDC42, FZD4, SOS1, RUNX1, TGFBR2*	1.74×10^{-6}
Pancreatic cancer	*SMAD3, SMAD2, CDC42, ARHGEF6, TGFBR2*	6.17×10^{-6}

Table 3.3. Gene-enrichment analysis of the interactome.
Genes associated to AID and DPACs, and their interacting partners (Fig. 3.4) were enriched for several immune related pathways. We display the ten most significantly enriched pathways. Genes within each pathway that were also within our query set are listed. Displayed p-values are adjusted for multiple testing (Materials and Methods).

Dataset	SNP	FM_{02} adjusted	$FM_{F.comb}$ adjusted	$FM_{S.comb}$ adjusted
Vitiligo GWAS1	rs2007899	$8.24x10^{-02}$	$1.42x10^{-01}$	$2.90x10^{-02}$
	rs12852381	$3.72x10^{-02}$	$6.51x10^{-02}$	$1.34x10^{-02}$
	rs143231802	$2.58x10^{-02}$	$3.62x10^{-02}$	$6.47x10^{-03}$
	rs4271099	$2.99x10^{-02}$	$5.27x10^{-02}$	$1.09x10^{-02}$
	rs4335270	$6.52x10^{-02}$	$1.23x10^{-01}$	$2.67x10^{-02}$
	rs4480250	$6.52x10^{-02}$	$1.23x10^{-01}$	$2.67x10^{-02}$
	rs17258266	$4.90x10^{-02}$	$7.55x10^{-02}$	$1.38x10^{-02}$
	rs4300122	$8.14x10^{-02}$	$1.61x10^{-01}$	$3.62x10^{-02}$
	rs5957594	$3.77x10^{-02}$	$7.27x10^{-02}$	$1.46x10^{-02}$
	rs34320000	$7.87x10^{-02}$	$1.59x10^{-01}$	$3.24x10^{-02}$
	rs5957596	$1.14x10^{-01}$	$2.01x10^{-01}$	$3.92x10^{-02}$
	rs10217856	$4.06x10^{-02}$	$6.84x10^{-02}$	$1.31x10^{-02}$
	rs5956287	$2.71x10^{-01}$	$2.51x10^{-01}$	$4.63x10^{-02}$
	rs12834182	$2.71x10^{-01}$	$2.51x10^{-01}$	$4.63x10^{-02}$
	rs1121546	$1.27x10^{-02}$	$2.79x10^{-02}$	$6.42x10^{-03}$
	rs5957620	$2.17x10^{-02}$	$3.51x10^{-02}$	$6.95x10^{-03}$
	rs9887587	$2.06x10^{-02}$	$2.97x10^{-02}$	$5.71x10^{-03}$
	rs150986507	$1.66x10^{-02}$	$3.31x10^{-02}$	$6.63x10^{-03}$
	rs12839589	$5.00x10^{-02}$	$8.37x10^{-02}$	$1.73x10^{-02}$
	rs33977652	$2.56x10^{-01}$	$2.66x10^{-01}$	$4.86x10^{-02}$
	rs138347087	$2.26x10^{-01}$	$2.39x10^{-01}$	$4.38x10^{-02}$
	rs35046609	$2.91x10^{-02}$	$4.81x10^{-02}$	$9.59x10^{-03}$
	rs60669023	$2.83x10^{-01}$	$2.43x10^{-01}$	$4.50x10^{-02}$
	rs16996189	$2.97x10^{-01}$	$2.60x10^{-01}$	$4.82x10^{-02}$
	rs5957651	$2.98x10^{-02}$	$5.49x10^{-02}$	$1.11x10^{-02}$
	rs148797601	$6.88x10^{-03}$	$1.94x10^{-02}$	$4.33x10^{-03}$
	rs148097246	$7.20x10^{-03}$	$2.04x10^{-02}$	$4.58x10^{-03}$
WT2 AS	rs7057428	$2.04x10^{-02}$	1.0	1.0
	rs5977756	$7.21x10^{-03}$	$9.05x10^{-02}$	$2.78x10^{-01}$
WT2 UC	rs5916435	$1.80x10^{-03}$	$1.22x10^{-03}$	$2.19x10^{-04}$
	rs5973636	$1.15x10^{-01}$	$4.76x10^{-02}$	$8.74x10^{-03}$
	rs6610386	$3.72x10^{-02}$	$3.76x10^{-02}$	$2.24x10^{-02}$
	rs59269143	$2.36x10^{-02}$	$2.26x10^{-02}$	$1.19x10^{-02}$
	rs5963157	$2.73x10^{-02}$	$2.66x10^{-02}$	$1.47x10^{-02}$
	rs7060409	$7.16x10^{-03}$	$6.50x10^{-03}$	$2.64x10^{-03}$
	rs62626573	$1.81x10^{-02}$	$1.63x10^{-02}$	$4.70x10^{-03}$
	rs35764713	$6.96x10^{-07}$	$3.31x10^{-09}$	$7.72x10^{-10}$
	rs5969304	$7.58x10^{-05}$	1.0	1.0
	rs6643227	$3.05x10^{-03}$	1.0	1.0
	rs6655215	$2.04x10^{-04}$	$5.75x10^{-03}$	$1.03x10^{-01}$

	rs12008980	$8.74x10^{-05}$	$5.73x10^{-02}$	$5.85x10^{-01}$

Table 3.S1. Significant SNP associations.
This table lists all significant associations (adjusted $P < 0.05$) in either the FM_{02} or FM_{comb} test. P-values are Bonferroni adjusted for the number of SNPs tested as listed in Table 3.1.

$FM_{F.comb}$			
Dataset	**Gene symbol**	**Truncated tail p-value**	**Truncated product p-value**
ALS Finland	TAF7L	0.000389	0.0018
ALS Finland	MAGEE2	0.00028	0.0012
ALS Finland	NAP1L2	0.00091	0.00034
ALS Finland	TTC3P1	0.000859	0.0013
ALS Finland	ZDHHC15	0.000413	0.0089
CASP	NLGN4X	0.000887	0.0166
Celiac disease CIDR	CENPI	0.0029	0.000523
Vitiligo GWAS1	PPP1R3F	0.000114	0.000496
Vitiligo GWAS1	LINC00632	0.0057	0.000772
Vitiligo GWAS1	FOXP3	0.000698	0.0015
Vitiligo GWAS1	BEND2	0.0018	0.000079
Vitiligo GWAS1	CENPI	0.000155	0.0026
Vitiligo GWAS2	IL13RA2	0.0021	0.000758
Vitiligo GWAS2	MCF2	0.00017	0.000576
CD WT1	CD40LG	0.009	0.000322
CD WT1	LINC00892	0.0013	0.000088
T2D WT1	MAGEC1	0.0275	0.000181
UC WT2	CASK	0.000138	0.0215
UC WT2	PRPS1	0.000133	0.000194
UC WT2	PAGE2B	0.0039	0.000012
UC WT2	SPANXN5	0.00091	0.0013
MS WT2	MAGEE1	0.000706	0.0023
$FM_{S.Comb}$			
ALS Finland	TAF7L	0.000547	0.000644
ALS Finland	NAP1L2	0.00057	0.000115
ALS Finland	ITM2A	0.000843	0.000307
ALS Finland	CENPI	0.001271	0.000175
ALS Finland	TMEM35	0.002775	0.000345
CASP	MIR505	$<1x10^{-6}$	0.001932
CASP	DCX	0.000757	0.00608
Celiac CIDR	IQSEC2	0.00053	0.00071
CD WT1	Y RNA	$<1x10^{-6}$	0.000052
CD WT1	LINC00892	0.001739	0.000529
UC WT2	PRPS1	0.000005	0.000005
UC WT2	CASK	0.000157	0.021124
UC WT2	GPR82	0.000209	0.001885
UC WT2	GPR34	0.000262	0.000162
UC WT2	PAGE2B	0.000482	0.000002
UC WT2	NAP1L6	0.001192	0.000429
MS Case Control	RP11-265P11.2	0.00303	0.000855
Vitiligo GWAS1	PPP1R3F	0.000006	0.000076
Vitiligo GWAS1	FOXP3	0.000022	0.000149

Vitiligo GWAS1	XRCC6P5	0.000081	0.001846
Vitiligo GWAS1	HUWE1	0.000362	0.001298
Vitiligo GWAS1	GAGE12H	0.000634	0.000634
Vitiligo GWAS1	GAGE10	0.001848	0.000266
Vitiligo GWAS2	MCF2	0.000078	0.000131
Vitiligo GWAS2	IL13RA2	0.000942	0.000354
Vitiligo GWAS2	RBMXL3	0.002653	0.000321
T2D GENEVA	ZCCHC12	0.001209	0.000653
T2D GENEVA	SNORA35	0.002123	0.000454
T2D GENEVA	IL13RA1	0.00635	0.000859
T2D WT1	MAGEC1	0.026251	0.000068
T1D WT1	ARX	0.000192	0.000489
T1D WT1	SRPK3	0.000469	0.008982
T1D WT1	PLXNB3	0.000487	0.007309
T1D WT1	RNU6-98P	0.000803	0.001921

Table 3.S2. All genes with either truncated tail or truncated product p-values < 1x10-3 for the $FM_{F.comb}$ and the $FM_{S.comb}$ test.

Dataset	Gene symbol	Truncated tail p-value	Truncated product p-value
ALS Finland	TAF7L	0.000126	0.000332
ALS Finland	NAP1L2	0.000451	0.000038
ALS Finland	ITM2A	0.0021	0.00041
CASP	PGRMC1	$<1x10^{-6}$	0.0046
CASP	ATP11C	0.000011	0.0092
CASP	DCX	0.000752	0.0048
CASP	MIR505	$<1x10^{-6}$	0.0039
MS case control	FANCB	0.000052	0.0013
MS case control	RP11-265P11.2	0.0025	0.000423
Vitiligo GWAS1	PPP1R3F	0.000066	0.000139
Vitiligo GWAS1	HUWE1	0.000822	0.0027
Vitiligo GWAS1	LINC00632	0.0137	0.000453
Vitiligo GWAS1	FOXP3	0.000111	0.000276
Vitiligo GWAS1	GAGE10	0.0016	0.000403
Vitiligo GWAS1	CENPI	0.000217	0.001
Vitiligo GWAS1	MPC1L	$<1x10^{-6}$	$<1x10^{-6}$
Vitiligo GWAS1	NAA10	0.00087	0.0028
Vitiligo GWAS2	IL13RA2	0.001014	0.000526
Vitiligo GWAS2	MCF2	0.000224	0.000559
Vitiligo GWAS2	RBMXL3	0.0019	0.000418
GENEVA T2D	RP4-562J12.2	0.000489	0.0013
CD WT1	ARHGEF6	0.0017	0.000366
CD WT1	CD40LG	0.0123	0.000223
CD WT1	LINC00892	0.001572	0.000048
T1D WT1	SRPK3	0.000327	0.0071
T1D WT1	ARX	0.000837	0.000565
T1D WT1	RNU6-98P	0.000716	0.0018
T1D WT1	PLXNB3	0.000522	0.0076
T2D WT1	MAGEC1	0.0264	0.000534
T2D WT1	SASH3	$<1x10^{-6}$	$<1x10^{-6}$
T2D WT1	DUSP9	0.0022	0.000553
UC WT2	CASK	0.000357	0.0199
UC WT2	PRPS1	0.000003	0.00001
UC WT2	NAP1L6	0.001063	0.000057
UC WT2	PAGE2B	0.012	0.000072
UC WT2	GPR34	0.0011	0.00061

Table 3.S3. All genes with either truncated tail or truncated product p-values $< 1x10^{-3}$ for the FM_{02} test.

Dataset	Gene	p-value (tail, product)	Replication dataset	p-value (tail, product)	combined p-value (tail, product)
FM_{02}					
Vitiligo GWAS1	PPP1R3F	$6.60x10^{-5}$, $1.39x10^{-4}$	Vitiligo GWAS2	$8.10x10^{-3}$, $2.70x10^{-3}$	$8.26x10^{-6}$, $5.93x10^{-6}$
Vitiligo GWAS1	FOXP3	$1.11x10^{-4}$, $2.76x10^{-4}$	Vitiligo GWAS2	$5.60x10^{-3}$, $5.40x10^{-3}$	$9.50x10^{-6}$, $2.15x10^{-5}$
Vitiligo GWAS1	GAGE10	$1.60x10^{-3}$, $4.03x10^{-4}$	Vitiligo GWAS2	$2.80x10^{-3}$, $3.80x10^{-3}$	$5.97x10^{-5}$, $2.20x10^{-5}$
CD WT1	ARHGEF6	$1.70x10^{-3}$, $3.66x10^{-4}$	UC WT2	$2.30x10^{-3}$, $3.10x10^{-3}$	$5.26x10^{-5}$, $1.67x10^{-5}$
$FM_{F.comb}$					
Vitiligo GWAS1	PPP1R3F	$1.14x10^{-4}$, $4.96x10^{-4}$	Vitiligo GWAS2	$3.70x10^{-3}$, $5.80x10^{-3}$	$6.61x10^{-6}$, $3.96x10^{-5}$
$FM_{S.comb}$					
Vitiligo GWAS1	PPP1R3F	$6.0x10^{-6}$, $7.60x10^{-5}$	Vitiligo GWAS2	$4.80x10^{-3}$, $1.30x10^{-3}$	$5.29x10^{-7}$, $1.69x10^{-6}$
	GAGE12H	$6.34x10^{-4}$, $6.34x10^{-4}$	Vitiligo GWAS2	$4.60x10^{-3}$, $4.60x10^{-3}$	$4.01x10^{-5}$, $4.01x10^{-5}$
	GAGE10	$1.85x10^{-3}$, $2.66x10^{-4}$	Vitiligo GWAS2	$2.90x10^{-3}$, $2.80x10^{-3}$	$7.05x10^{-5}$, $1.13x10^{-5}$
Sex Difference					
CD WT1	C1GALT1C1	$1.97x10^{-3}$, $2.63x10^{-4}$	UC WT2	$1.39x10^{-2}$, $1.14x10^{-2}$	$3.15x10^{-4}$, $4.11x10^{-5}$

Table 3.S4. Gene-based associations replicating in similar diseases.
Table of genes with nominal $P < 1x10^{-3}$ that replicated in a dataset of the same or similar disease. Combined p-values were calculated using Fisher's method.

Dataset	p-value (tail, product)
ALS Finland	$1.10x10^{-2}$, $1.00x10^{-3}$
ALS Irish	$2.70x10^{-2}$, $1.60x10^{-2}$
CASP	0.91, 0.64
CIDR Celiac	$2.9x10^{-3}$, $5.23x10^{-4}$
NIDDK CD	0.17, 0.16
MS case control	0.91, 0.38
Vitiligo GWAS1	$1.55x10^{-4}$, $2.6x10^{-3}$
Vitiligo GWAS2	0.827, 0.65
Geneva T2D	0.17, 0.19
WT1 CD	0.83, 0.20
WT1 T1D	0.85, 0.49
WT1 RA	0.83, 0.29
WT1 T2D	0.93, 0.54
WT2 UC	0.88, 0.45
WT2 MS	0.81, 0.67
WT2 AS	0.11, 0.11

Table 3.S5. CENPI association p-values for the $FM_{F.comb}$ test across the 16 datasets.

Dataset	Gene	p-value (tail, product)	Alternate dataset	p-value (tail, product)	combined p-value (tail, product)
FM_{02}					
ALS Finland	NAP1L2	$4.51x10^{-4}$, $3.80x10^{-5}$	UC WT2	$5.70x10^{-3}$, $3.70x10^{-3}$	$3.57x10^{-5}$, $2.36x10^{-6}$
			Vitiligo GWAS1	$1.0x10^{-2}$, $1.40x10^{-2}$	$6.00x10^{-5}$, $8.22x10^{-6}$
ALS Finland	ITM2A	$2.10x10^{-3}$, $4.10x10^{-4}$	Celiac Disease CIDR	$7.90x10^{-3}$, $1.06x10^{-2}$	$1.99x10^{-4}$, $5.80x10^{-5}$
MS case control	FANCB	$5.20x10^{-5}$, $1.30x10^{-3}$	RA WT1	$3.80x10^{-3}$, $1.10x10^{-2}$	$3.25x10^{-6}$, $1.74x10^{-4}$
Vitiligo GWAS1	CENPI	$2.17x10^{-4}$, $1.00x10^{-3}$	ALS Finland	$2.40x10^{-3}$, $2.00x10^{-3}$	$8.06x10^{-6}$, $2.82x10^{-5}$
T2D GENEVA	RP4-562J12.2	$4.89x10^{-4}$, $1.30x10^{-4}$	CD NIDDK	$3.41x10^{-2}$, $3.93x10^{-2}$	$2.00x10^{-4}$, $5.56x10^{-4}$
			WT2 AS	$5.60x10^{-2}$, $4.30x10^{-2}$	$3.15x10^{-4}$, $7.32x10^{-5}$
T2D WT1	MAGEC1	$2.64x10^{-2}$, $5.34x10^{-4}$	MS case control	$6.70x10^{-3}$, $8.50x10^{-3}$	$1.71x10^{-3}$, $6.04x10^{-5}$
UC WT21	NAP1L6	$1.06x10^{-3}$, $5.70x10^{-5}$	ALS Finland	$3.10x10^{-3}$, $5.50x10^{-3}$	$4.49x10^{-5}$, $5.01x10^{-6}$
$FM_{F.comb}$					
CASP	NLGN4X	$8.87x10^{-4}$, $1.66x10^{-2}$	Vitiligo GWAS2	$1.21x10^{-2}$, $1.31x10^{-2}$	$1.34x10^{-4}$, $2.05x10^{-3}$
			CIDR Celiac Disease	$5.10x10^{-2}$, $4.90x10^{-2}$	$4.98x10^{-4}$, $6.66x10^{-3}$
Vitiligo GWAS1	BEND2	$1.80x10^{-3}$, $7.90x10^{-5}$	T2D WT1	$9.30x20^{-3}$, $1.29x10^{-2}$	$2.01x10^{-4}$, $1.51x10^{-5}$
Vitiligo GWAS1	CENPI	$1.55x10^{-4}$, $2.60x10^{-3}$	ALS Finland	$1.12x10^{-2}$, $1.00x10^{-3}$	$2.48x10^{-5}$, $3.60x10^{-5}$
			Celiac CIDR	$2.90x10^{-3}$, $5.23x10^{-4}$	$7.02x10^{-6}$, $1.97x10^{-5}$
Vitiligo GWAS2	MCF2	$1.70x10^{-4}$, $5.76x10^{-4}$	MS WT2	$2.31x10^{-2}$, $2.50x10^{-2}$	$5.28x10^{-5}$, $1.75x10^{-4}$
CD WT1	LINC00892	$1.30x10^{-3}$, $8.80x10^{-5}$	MS WT2	$2.42x10^{-2}$, $1.99x10^{-2}$	$3.58x10^{-4}$, $2.50x10^{-5}$
T2D WT1	MAGEC1	$2.75x10^{-2}$, $1.81x10^{-4}$	MS case control	$1.42x10^{-2}$, $1.50x10^{-2}$	$3.46x10^{-3}$, $3.75x10^{-5}$
MS WT2	MAGEE1	$7.06x10^{-4}$, $2.30x10^{-3}$	ALS Finland	$3.23x10^{-2}$, $2.36x10^{-2}$	$2.67x10^{-4}$, $5.87x10^{-4}$
$FM_{S.comb}$					
ALS Finland	NAP1L2	$5.7x10^{-4}$, $1.15x10^{-4}$	UC WT2	$8.30x10^{-3}$, $7.1x10^{-3}$	$6.27x10^{-5}$, $1.23x10^{-5}$
	ITM2A	$8.43x10^{-4}$, $3.07x10^{-4}$	Celiac CIDR	$6.5x10^{-3}$, $1.13x10^{-2}$	$7.19x10^{-5}$, $4.71x10^{-5}$
	CENPI	$1.27x10^{-3}$, $1.75x10^{-4}$	Vitiligo GWAS1	$1.60x10^{-3}$, $5.90x10^{-3}$	$2.89x10^{-5}$, $1.53x10^{-5}$
	TMEM35	$2.78x10^{-3}$, $3.45x10^{-4}$	Vitiligo GWAS1	$3.80x10^{-3}$, $6.20x10^{-3}$	$1.31x10^{-4}$, $3.01x10^{-5}$
CD WT1	LINC00892	$1.73x10^{-3}$, $5.29x10^{-4}$	MS WT2	$6.30x10^{-3}$, $6.40x10^{-3}$	$1.35x10^{-4}$, $4.60x10^{-5}$
			Vitiligo GWAS1	$2.30x10^{-2}$, $2.89x10^{-2}$	$4.41x10^{-4}$, $1.85x10^{-4}$
UC WT2	GPR34	$2.62x10^{-4}$, $1.62x10^{-4}$	MS WT2	$5.60x10^{-3}$, $1.10x10^{-2}$	$2.12x10^{-5}$, $2.54x10^{-5}$
	NAP1L6	$1.19x10^{-3}$, $4.29x10^{-4}$	ALS Finland	$4.00x10^{-3}$, $1.06x10^{-2}$	$6.31x10^{-5}$, $6.05x10^{-5}$

MS case control	RP11-265P11.2	3.03×10^{-3}, 8.55×10^{-4}	T2D WT1	4.42×10^{-2}, 4.68×10^{-2}	1.32×10^{-3}, 4.45×10^{-4}
T2D GENEVA	SNORA35	2.12×10^{-3}, 4.54×10^{-4}	AS WT2	2.40×10^{-3}, 6.70×10^{-3}	6.71×10^{-5}, 4.17×10^{-5}
	IL13RA1	6.35×10^{-3}, 8.59×10^{-4}	AS WT2	6.20×10^{-3}, 7.20×10^{-3}	4.39×10^{-4}, 8.04×10^{-5}
T2D WT1	MAGEC1	2.63×10^{-2}, 6.80×10^{-5}	MS case control	1.00×10^{-2}, 1.54×10^{-2}	2.43×10^{-3}, 1.55×10^{-5}
Sex difference					
ALS Finland	MAGEE2	6.5×10^{-4}, 1.94×10^{-3}	Vitiligo GWAS1	3.08×10^{-2}, 1.64×10^{-2}	2.37×10^{-4}, 3.61×10^{-4}
	NDP	1.41×10^{-3}, 9.34×10^{-4}	CD WT1	8.60×10^{-3}, 1.33×10^{-2}	1.49×10^{-4}, 1.53×10^{-4}
CASP	NLGN4X	2.34×10^{-4}, 1.65×10^{-2}	Vitiligo GWAS1	4.52×10^{-2}, 4.33×10^{-2}	1.32×10^{-4}, 5.89×10^{-3}
Celiac CIDR	CENPI	4.4×10^{-3}, 2.08×10^{-4}	ALS Finland	2.03×10^{-2}, 1.78×10^{-2}	9.22×10^{-4}, 5.00×10^{-5}
			ALS Irish	9.80×10^{-3}, 4.40×10^{-3}	4.88×10^{-4}, 1.36×10^{-5}
Vitiligo GWAS1	BEND2	3.99×10^{-3}, 1.28×10^{-4}	MS case control	4.60×10^{-2}, 5.20×10^{-2}	1.76×10^{-3}, 8.60×10^{-5}
Vitiligo GWAS2	MCF2	7.00×10^{-4}, 1.93×10^{-3}	MS WT2	2.38×10^{-2}, 2.12×10^{-2}	2.00×10^{-4}, 4.54×10^{-4}
T2D GENEVA	EFHC2	6.09×10^{-4}, 1.12×10^{-3}	RA WT1	1.58×10^{-2}, 1.40×10^{-3}	1.21×10^{-4}, 2.42×10^{-5}
RA WT1	MIR320D2	8.69×10^{-3}, 5.68×10^{-4}	ALS Irish	2.39×10^{-2}, 2.64×10^{-2}	1.97×10^{-3}, 1.82×10^{-4}

Table 3.S6. Gene-based associations replicating in other diseases.
This table lists genes with nominal $P < 1\times10^{-3}$ that replicated in a disease of a different phenotype (Methods). Combined p-values were calculated using Fisher's method. *We assumed a p-value = 1×10^{-6} in the truncated tail p-value for WT CD when calculating the combined p-value.

Dataset	Gene symbol	Truncated tail p-value	Truncated product p-value
ALS Finland	MAGEE2	6.50×10^{-4}	1.93×10^{-3}
ALS Finland	NDP	1.41×10^{-3}	9.34×10^{-4}
CASP	NLGN4X	2.34×10^{-4}	0.017
CIDR Celiac	CENPI	4.42×10^{-3}	2.08×10^{-4}
CD WT1	C1GALT1C1	1.97×10^{-3}	2.63×10^{-4}
UC WT2	SPANXN5	2.72×10^{-4}	3.45×10^{-4}
UC WT2	XAGE5	2.21×10^{-3}	3.45×10^{-4}
MS case control	ZNF449	7.22×10^{-4}	3.11×10^{-3}
MS case control	BMX	9.91×10^{-4}	2.49×10^{-3}
Vitiligo GWAS1	BEND2	3.99×10^{-3}	1.28×10^{-4}
Vitiligo GWAS2	MCF2	7.00×10^{-4}	1.93×10^{-3}
T2D GENEVA	EFHC2	6.09×10^{-4}	1.1×10^{-3}
T2D WT1	SASH3	$<1 \times 10^{-6}$	$<1 \times 10^{-6}$
T2D WT1	CSTF2	1.63×10^{-3}	8.17×10^{-4}
T2D WT1	SNORA9	1.63×10^{-3}	8.43×10^{-4}
T2D WT1	SYTL4	2.27×10^{-3}	3.27×10^{-4}
RA WT1	MIR320D2	8.69×10^{-3}	5.68×10^{-4}

Table 3.S7. All genes with either the truncated tail or truncated product p-values $< 1 \times 10^{-3}$ for the sex difference test.

$Gene_1$	$Gene_2$	r	P-value
ARHGEF6	*EFHC2*	0.208	$4.44x10^{-10}$
ARHGEF6	*IL13RA1*	0.267	$7.46x10^{-16}$
ARHGEF6	*PPP1R3F*	0.302	$4.41x10^{-20}$
BEND2	*EFHC2*	0.185	$3.39x10^{-8}$
C1GALT1C1	*EFHC2*	0.449	$6.80x10^{-45}$
C1GALT1C1	*FANCB*	0.178	$1.12x10^{-7}$
C1GALT1C1	*IL13RA1*	0.221	$3.31x10^{-11}$
C1GALT1C1	*ITM2A*	0.226	$1.27x10^{-11}$
C1GALT1C1	*PPP1R3F*	0.278	$4.57x10^{-17}$
EFHC2	*FANCB*	0.192	$8.49x10^{-9}$
EFHC2	*IL13RA1*	0.2	$2.17x10^{-9}$
EFHC2	*ITM2A*	0.291	$1.13x10^{-18}$
EFHC2	*PPP1R3F*	0.496	$6.03x10^{-56}$
FANCB	*PPP1R3F*	0.183	$4.33x10^{-8}$
ITM2A	*PPP1R3F*	0.276	$7.89x10^{-17}$
NLGN4X	*TMEM35*	0.193	$8.31x10^{-9}$

Table 3.S8. Pairs of X-linked genes that are significantly co-expressed.
We assayed whether associated X-linked genes were significantly co-expressed in samples of healthy individuals (see methods in main text). "r" denotes the spearman's rank correlation coefficient, with the p-value listed in the adjacent column ("P-value").

Gene symbol
OTUD5
TLR8
CFP
RNF128
PRKX
APLN
BTK
IL3RA
IKBKG
IRAK1
CD40LG
SH2D1A
XIAP
NOX1
CXCR3
IL2RG
EDA
FOXP3
WAS
CYBB
TAB3
TLR7
CD99
DDX3X
CSF2RA
IL9R
BCAP31

Table 3.S9. List of genes in the KEGG/GO immune gene set.

WT2 UC, WT1 CD

			Original p-value	
Basepair	SNP	meta-analysis p-value	WT2 UC	WT1 CD
135670593	rs2518886	0.00005669	0.005774	0.00003564
135670824	rs12852548	0.00008458	0.008264	0.0000392
135671221	rs2807259	0.0000702	0.007347	0.00003564
135673225	rs5930965	0.00006327	0.00631	0.00003968
135673610	rs7890404	0.0000173	0.003203	0.00001065
135674044	rs2807260	0.00004705	0.005757	0.00002743
135674388	rs2518888	0.00007253	0.007565	0.00003758
135674543	rs2518889	0.00007253	0.007565	0.00003758
135675755	rs2518891	0.00006978	0.007508	0.00003558
135676896	rs2518892	0.00007248	0.007453	0.00003758
135677559	rs2518893	0.00008051	0.008385	0.00003758
135677710	rs2518894	0.00006895	0.006859	0.00003758
135678840	rs2518895	0.00007248	0.007453	0.00003758
135679053	rs2518896	0.00007248	0.007453	0.00003758
135679059	rs2518897	0.00007248	0.007453	0.00003758
135679631	rs2518899	0.00007248	0.007453	0.00003758
135679960	rs2518900	0.00007248	0.007453	0.00003758
135680078	rs2518901	0.00006039	0.00753	0.00002765
135680774	rs2518902	0.00007248	0.007453	0.00003758
135680970	rs12556398	0.00007972	0.007776	0.00004124
135681823	rs2518904	0.00007611	0.007371	0.00004155
135681929	rs73242348	0.00008436	0.008296	0.00004057
135682887	rs12007112	0.00003256	0.002824	0.00004155
135683500	rs12012314	0.00007501	0.007918	0.00003735
135683508	rs73228703	0.00007501	0.007918	0.00003735
135684162	rs12848318	0.00007801	0.007726	0.00003891
135685169	rs12559890	0.00007801	0.007726	0.00003891
135685563	rs12014670	0.00007801	0.007726	0.00003891
135687540	rs12559116	0.0000454	0.005664	0.00002597
135689560	rs12558063	0.00006444	0.01388	0.00001018

Table 3.S10a. All SNPs with a meta-analysis p-value < $1x10^{-4}$ for the IBD disease set.

CASP, Vitiligo GWAS1

			Original p-value	
BP	SNP	meta-analysis p-value	Vitiligo GWAS1	CASP
13539543	rs11797576	0.00007132	0.00001217	0.113
120356233	rs12852381	0.00003984	2.358x10-07	0.7948
120361243	rs143231802	0.00005031	1.635x10-07	0.5694
120363815	rs4271099	0.00002333	1.899x10-07	0.7179
120363833	rs4335270	0.0000365	4.137x10-07	0.7179
120364808	rs4480250	0.00003571	4.137x10-07	0.7179
120372902	rs17258266	0.00004401	3.107x10-07	0.4317
120377156	rs4300122	0.0000176	5.165x10-07	0.384
120377239	rs5957593	0.00001991	6.505x10-07	0.3878
120377928	rs5957594	0.00002138	2.392x10-07	0.4953
120384209	rs34320000	0.00001928	4.989x10-07	0.3873
120385628	rs5957596	0.00002515	7.216x10-07	0.3929
120397795	rs10217856	0.00001654	2.575x10-07	0.4652
120399965	rs188743539	0.00001312	0.000001017	0.2152
120402522	rs5956287	0.00005116	0.00000172	0.5429
120403437	rs12834182	0.00005116	0.00000172	0.5429
120412794	rs1121546	0.00000699	8.043x10-08	0.4616
120414055	rs5957620	0.000007265	1.376x10-07	0.408
120417125	rs9887587	0.000006633	1.304x10-07	0.381
120422623	rs150986507	0.00001451	1.051x10-07	0.3113
120423853	rs12839589	0.00001188	3.173x10-07	0.3853
120428148	rs33977652	0.00004402	0.000001621	0.4982
120432609	rs138347087	0.00009268	0.000001435	0.4011
120440791	rs35046609	0.00001107	1.843x10-07	0.423
120449026	rs60669023	0.000059	0.000001795	0.5376
120456282	rs16996189	0.00006177	0.000001882	0.5376
120457727	rs5957651	0.00001094	1.892x10-07	0.4186
120506286	rs148797601	0.00001516	4.363x10-08	0.3956
120514199	rs148097246	0.00001568	4.568x10-08	0.3956
120581955	rs139713212	0.00006601	0.000003853	0.1313
120615789	rs12387331	0.00006826	0.000005681	0.2252
120671167	rs111852695	0.00003387	8.647x10-07	0.1378
120706195	rs140636073	0.00006317	0.000003378	0.1218

Table 3.S10b. All SNPs with a meta-analysis p-value < $1x10^{-4}$ for the skin-related disease set.

Classic Autoimmune (CASP, CIDR celiac disease, WT2 AS, WT2 MS, WT2 UC, WT1 CD, WT1 RA, WT1 T1D)										
			Original p-values							
BP	SNP	meta-analysis p-value	CASP	CIDR	AS	MS	UC	CD	RA	T1D
135867861	rs6635322	0.00007194	0.1086	0.2043	0.08659	0.06399	0.001849	0.0007687	0.05755	0.3813
135876353	rs5930994	0.00009651	0.1087	0.1314	0.06854	0.05932	0.003505	0.0003842	0.09112	0.3001
135879166	rs5930995	0.00005965	0.1093	0.1251	0.06154	0.07357	0.002276	0.0009112	0.03484	0.2871
135883888	rs5930998	0.00006352	0.1135	0.1161	0.06302	0.07474	0.00133	0.0009793	0.04361	0.2696

Table 3.S10c. All SNPs with a meta-analysis p-value < $1x10^{-4}$ for the classic autoimmune disease set

Chapter 4 Principal component analysis characterizes shared pathogenetics from genome-wide association studies

(Chang and Keinan 2014)

4.1 Introduction

Comorbidity studies show that some distinct diseases tend to co-occur in the same individuals (Sowers 1998; Broadley, Deans et al. 2000; Somers, Thomas et al. 2009; Zaccara 2009; Marrie, Horwitz et al. 2011; Sardu, Cocco et al. 2012), pointing to a shared genetic and/or environmental component. In the era of genome-wide association studies (GWASs), direct evidence of shared genetic risk factors of diseases comes to light (Solovieff, Cotsapas et al. 2013). For example, while it has been previously shown that rheumatoid arthritis and type-1 diabetes co-occur (Somers, Thomas et al. 2009), GWASs have identified 12 genes associated with both diseases (Hakonarson, Grant et al. 2007; ⁋The Wellcome Trust Case Control Consortium 2007; Todd, Walker et al. 2007; Barrett, Clayton et al. 2009; Hindorff, Sethupathy et al. 2009; Stahl, Raychaudhuri et al. 2010; Festen, Goyette et al. 2011; Okada, Terao et al. 2012; Hindorff, MacArthur et al. 2013). More broadly, disease genes obtained from the Online Mendelian Inheritance in Man (Hamosh, Scott et al. 2005) were used to assemble the Human Disease Network (HDN) (Goh, Cusick et al. 2007; Darabos, Desai et al. 2013), a visual representation of genetic similarity between diseases. Pleiotropy of complex diseases and traits has also been explored by searching genome-wide for variants implicated in more than one disease (Festen, Goyette et al. 2011; Zhernakova, Stahl et al. 2011; Ellinghaus, Ellinghaus et al. 2012). Such

studies promise to reveal shared genes and offer an expanded understanding from a genetic standpoint of why some diseases tend to co-occur.

Methods for exploring shared genetic risk variants between diseases belong to two main categories and have been recently reviewed (Solovieff, Cotsapas et al. 2013). In the first category of methods, variants are tested for association to a pair or more of diseases being investigated. In one set of methods, a GWAS is carried out on individuals with different diseases pooled together (‚The Wellcome Trust Case Control Consortium 2007; Festen, Goyette et al. 2011; Zhernakova, Stahl et al. 2011; Ellinghaus, Ellinghaus et al. 2012) or by analyzing information for multiple diseases available for the same individuals (Lee, Bergen et al. 2011; Hartley, Monti et al. 2012). Alternatively, and based only on summary statistics of the association test for each single nucleotide polymorphism (SNP), one can simply combine p-values from several GWASs using Fisher's method (Fisher 1925). The CPMA (cross-phenotype meta-analysis) statistic (Cotsapas, Voight et al. 2011) is another statistic that tests whether a SNP is associated to more than one phenotype. In addition, methods such as the conditional false discovery rate or mixed-models for multiple traits have used known pleiotropy between diseases or traits to increase power (Korte, Vilhjalmsson et al. 2012; Andreassen, Thompson et al. 2013). Studies employing these methods have found shared associations between pairs of diseases such as Crohn's disease and celiac disease (Festen, Goyette et al. 2011), other autoimmune disease pairs (Zhernakova, Stahl et al. 2011; Ellinghaus, Ellinghaus et al. 2012), bipolar and schizophrenia (Andreassen, Thompson et al. 2013) and multiple sclerosis and schizophrenia (Andreassen, Harbo et al. 2014). They have additionally shown that SNPs associated with one autoimmune disease are likely to be associated to other (though not all) autoimmune phenotypes (Cotsapas, Voight et al. 2011).

The second category of methods focuses on using shared variants to learn about the genetic similarity between diseases. One method employed by Sirota *et al.* utilizes the correlation between association signals across many SNPs to assess the similarity between pairs of diseases and showed that there are likely two distinct autoimmune classes where a risk allele for one class may be protective in another (Sirota, Schaub et al. 2009). While another method uses a classifier approach to identify diseases that are similar (Schaub, Kaplow et al. 2009). A linear mixed model approach can also be applied to assess the shared genetic variation between two diseases (Korte, Vilhjalmsson et al. 2012; Lee, Ripke et al. 2013).

These exciting new methods are powerful for studying shared genetic risk variants between diseases. At the same time, overcoming some of their limitations can improve the study of shared pathogenesis using data from multiple GWASs. First, some methods have focused on analysis of individual SNPs. Though this is well suited for scenarios of a single causal SNP in a locus, they would lose power when several causal SNPs exist or if different SNPs tag the same underlying causal variant, which is especially relevant for diseases with rare causal variants (Wang, Dickson et al. 2010; Chang and Keinan 2012) and when the different GWASs are across different populations (Marigorta and Navarro 2013) or have used different genotyping arrays. Second, in one study where the correlation between association statistics of different studies is used to determine shared disease etiology, the correlation statistic weighs all variants equally, whether or not they play a role in disease susceptibility (Sirota, Schaub et al. 2009). Third, most methods assume as known which diseases share pathogenesis, and while the shared pathogenesis of autoimmune disease has been well established (Sirota, Schaub et al. 2009; Cotsapas, Voight et al.

2011), it is worthwhile to study shared pathogenesis of other disease classes (Yancik, Havlik et al. 1996; McElroy 2004; Zaccara 2009). And fourth, while some approaches perform well for two correlated traits or diseases, extending the analysis to more than two traits can become difficult (Korte, Vilhjalmsson et al. 2012).

In this study, we present a novel method, *disPCA*, which uses principal component analysis (PCA) to learn about the shared genetic risk of distinct diseases. PCA maps data from the original axes into new axes in principal component (PC) space via a stretch and rotation of the original axes. Each new axis or PC captures the maximal level of variation in the data not captured by previous PCs. Thus, each PC can potentially tell a different, orthogonal story regarding the data. Our method is based on summary level statistics from GWASs of different diseases. We combine data from individual SNPs into gene-based statistics via several p-value combination methods. PCA is applied to a matrix across genes and GWAS datasets, with entries representing the strength of association (p-value) between a gene and the disease studied in a dataset. This method is gene-centric, with the PCA weighing genes by their role in differentiating between different GWAS, and can be applied to study multiple diseases without prior knowledge of their shared pathogenesis, thereby overcoming all the limitations of existing methods outlined above. *disPCA* also accounts for potential confounders due to methodological differences between studies, such as in genotyping array, which can otherwise lead to these differences being captured by the PCA.

Equipped with this novel method and with data from 31 GWAS datasets, we considered the level of shared pathogenesis between diseases and classes of diseases from all genes, which we term

shared pathogenetics. Diseases with more similar underlying genetics are more likely to be located closer together in PC space. As PCA is a non-parametric method, it makes no assumptions regarding which diseases are more similar and does not aim to model it, thereby allowing discovery of new relationships between diseases by examining the top PCs. Each PC is a linear combination of genes, with the leading PCs expected to give more weight to genes that distinguish well between diseases. Diseases with no separation along any PC indicate that they tend to share the pathogenetics underlying that PC. By studying the set of genes underlying a PC for enrichment in specific pathways, we can further assess the function and relationship of genes that separate different disease clusters in PC space.

4.2 Materials and Methods

4.2.1 disPCA

We developed a method, *disPCA,* for studying the relationship between diseases based on their level of disease risk genes shared. The method works on the gene-level by first combining information from all SNPs in and around each gene. Considering gene-level statistics compensates for different tag SNPs being associated in different datasets even in cases where they capture the same causal variant. It also aggregates information across multiple tag SNPs in each dataset, as well as allows for different underlying causal variants in the same gene being associated with the risk of different diseases. To be widely applicable, *disPCA* is based solely on the p-values of association of each SNP with the disease under study. Importantly, all SNPs and consequently all genes are considered, rather than focusing on genes that meet a genome-wide significance level of association with a disease. We apply PCA to many different GWASs to axiomatically find and assign importance to genes based on their contribution to distinguishing between diseases and disease classes. The ensuing distance between different disease datasets in PC space inversely corresponds to their level of shared pathogenetics.

4.2.2 Gene-level significance levels

For each protein-coding gene from the HGNC database (Gray, Daugherty et al. 2013), we mapped all SNPs that are in the gene or within 0.01cM from it (genetic distances were determined via the Oxford genetic map based on HapMap2 data (Myers, Bottolo et al. 2005; Frazer, Ballinger et al. 2007)). We discarded all SNPs that were not mapped to within 0.01cM of

any gene. If a SNP lay between two genes, it was assigned to the closer gene. For each GWAS dataset, we determined the significance of association of each gene with the assayed disease using the following simulation procedure. Let the observed p-value of a gene be the minimum p-value of the *n* SNPs mapped to the gene. We compared the observed p-value to that of 100,000 groups of *n* consecutive SNPs chosen in random. Based on these groups, we assign a new p-value to each gene as the proportion of groups for which the observed minimum p-value for that gene is less significant than that of the group. This random sampling procedure may be biased in regions of high linkage disequilibrium (LD) when mapping SNPs to genes using genetic distance (e.g. consecutive SNPs in regions of high LD will be more correlated than those in regions of lower LD). However, for any given gene, these will equally affect each of the datasets. To validate this, we also applied *disPCA* to p-values obtained from mapping SNPs to genes using physical distance: a SNP was mapped to a gene if it was in the gene or within 10kb of it. Comparing these results to results based on mapping via genetic coordinates revealed the same clustering of diseases (Figure 4.S1). Furthermore, average loading of genes with the top 50 loadings on the first two PCs were significantly correlated ($r>0.67$, p-value $< 8.4 \times 10^{-8}$, Table 4.S1). Thus, in the main text we present results based on mapping by genetic distance as described above.

To consider information from beyond only the most significant SNP in a gene, we also implemented truncated tail strength (Jiang, Zhang et al. 2011) and truncated product (Zaykin, Zhivotovsky et al. 2002) to combine p-values in each gene in replacement of the minimum p-value, and followed a similar procedure for assigning new gene-level p-values. For the analyses presented in the following, results from all methods were similar though results with the

minimum p-value approach clusters similar diseases better (Figure 4.S2-S3). We thus only report in the main text results from the minimum p-value approach. Code to carry out this procedure is publicly available at http://keinanlab.cb.bscb.cornell.edu/content/tools-data.

4.2.3 PCA implementation and confounders

Assume a matrix Z, a d x g matrix of the $-\log_{10}$ gene-level p-values, where d is the number of GWAS datasets, and g is the number of genes present in all datasets. We center the matrix by subtracting the column means from each column. Thus the centered matrix B has entries:

$$B_{i,j} = Z_{i,j} - \frac{\sum_{k=1}^{d} Z_{k,j}}{d} \qquad (1)$$

To obtain the PCs of matrix B, we must find the eigenvectors and eigenvalues of its covariance matrix BB^T. Let v_i be a vector of length d and let λ_i be a scalar. v_i is the eigenvector and λ the eigenvalue of BB^{T} if the following is satisfied:

$$(BB^T)v_i = \lambda_i v_i \qquad (2)$$

The principal components of B are the normalized eigenvectors of its covariance matrix, BB^T, where the eigenvectors are ordered such that the largest eigenvalue corresponds to the first principal component. Each eigenvector is additionally orthogonal to all other eigenvectors. Thus, from (2), we can decompose BB^T as follows:

$$BB^T = U \sum U^T \qquad (3)$$

Where the columns of U contain the principal components and $\sum$ is a diagonal matrix with entries equal to the eigenvalues of B's covariance matrix. One can similarly construct the singular value decomposition (SVD) of B. The SVD of B can be written as:

$$B = VDW^T \qquad (4)$$

where V is a d x d matrix, D is a d x g diagonal matrix, and W is a g x g matrix. V and W contain the left and right singular vectors of B, respectively, and D contains the singular values of B in its diagonal. Substituting equation (4) for B in equation (3), we find that

$$BB^T = (VDW^T)(WDV^T) = VD^2V^T = U\sum U^T \qquad (5)$$

Thus, the principal components of B, the eigenvectors of its covariance matrix, are equivalent to the left singular vectors of B. In addition, the eigenvalues of B are equivalent to the square of its singular values.

We applied SVD to the matrix B using the R (R Core Team 2013) implementation of PCA/SVD (*prcomp*), with no scaling of the data. Due to the heterogeneity of the GWAS datasets (Table 4.S2), variation uncovered by PCA can also reflect differences in features such as genotyping array, association method, and sample size, rather than underlying disease risk genes. To ensure that these features did not influence our results, we first tested each gene for association with each of these features. Let $z_i = Z_i$, be the vector corresponding to the association statistic for gene i across the d datasets. We considered a linear regression of z_i as a function of the covariates: $z_i = \alpha + b_{i,1}C_1 + b_{i,2}C_2 + b_{i,3}C_3 + \varepsilon$, where C_1, C_2, C_3 are vectors of length d that represent the genotyping array, association method and the $\log_{10}$ of the sample size respectively, in each of the studies (Table 4.S2). Testing the significance of regression coefficients can reveal genes that are associated with any of these potential confounders. In our following analysis, 19 genes were significantly associated with association method. However, genes not significantly associated to the above confounders may similarly have an effect. Hence, we also applied SVD (as described above) to the residualized matrix, namely matrix R with rows

$R_{i,\bullet} = z_i - (\alpha + b_{i,1}C_1 + b_{i,2}C_2 + b_{i,3}C_3)$. We found that applying SVD to *R* results in the top PCs capturing a higher fraction of the variance of the data than when applied to the original matrix *Z*, though results are qualitatively similar between the two. We thus present results derived from the residualized matrix *R*. Resulting distances between datasets were assessed visually by plotting datasets in PC space. To quantify the clustering of datasets, we additionally applied hierarchical clustering in R (R Core Team 2013) (*hclust*) to the Euclidean distance between pairs of datasets across the first two PCs.

4.2.4 Simulation study

We simulated a matrix *Z* for two disease classes, each with 5 diseases (A_1,A_2,A_3,A_4,A_5,B_1,B_2,B_3,B_4,B_5) and 10,000 genes. In general, under the null hypothesis of a region containing no risk variant and assuming no confounding factors (e.g. population stratification), p-values should be uniformly distributed between 0 and 1. On the other hand, associated risk variants should be enriched for smaller p-values. We thus considered three sets of genes. The p-values for the first set of genes was drawn from the *U*(0,1) distribution for all diseases, thus no pleiotropy was captured in this set of genes. The second set of genes was distributed *U*(0,0.05) for the first disease class ($A_1,\ldots,A_5$) and distributed *U*(0,1) for the second disease class ($B_1,\ldots,B_5$). Finally the third set of genes was distributed *U*(0,0.05) for the following diseases: A_1, A_2, B_1, B_2 and distributed *U*(0,1) for all other diseases. Thus the second set of genes simulates pleiotropy between diseases in disease class A, while the last set of genes simulates pleiotropy between diseases in both disease classes.

4.2.5 Disease and pathway enrichment analysis

Disease enrichment analysis was completed using the online tool WebGestalt (Zhang, Kirov et al. 2005; Wang, Duncan et al. 2013) to query the PharmGKB (Whirl-Carrillo, McDonagh et al. 2012) database. WebGestalt tests for enrichment of a category of genes in the observed set of genes using the hypergeometric test (Zhang, Kirov et al. 2005). Bonferroni correction for multiple tests was applied and all reported p-values are following this correction. We restricted analysis to categories that contained a minimum of 5 genes in our analysis with the largest 50 weightings in the top two PCs. For gene categories with overlapping or the same set of genes, we list the most significant category. To reduce biases introduced by the clustering of genes with similar function, we filtered our list of genes with the top 50 loadings on the top two PCs by removing the latter gene out of a pair of genes within 0.1cM of each other. We then applied WebGestalt to this filtered subset of genes.

Pathway enrichment analysis was completed using the Gene Set Enrichment Analysis (GSEA) tool (Subramanian, Tamayo et al. 2005). GSEA sorts genes according to a score, which here is the weighting of a gene in the PC under study. It then assesses whether genes belonging to a certain category (e.g. pathway) are non-randomly distributed in the sorted list. As input to GSEA, we utilized the weights of genes in the top two PCs. GSEA carried out 10,000 gene-set permutations to determine FDR (false discovery rate) q-values. We queried the BioCarta and KEGG pathway databases. We restricted analysis to categories that contained a minimum of 5 genes in our analysis. Throughout we present enrichment analysis only for the top two PCs, though other PCs are available and can be assayed for further insight into the diseases studied. As above, to reduce biases introduced by the clustering of genes with similar function, we

filtered our full list of genes by removing the latter gene out of a pair of genes within 0.1cM of each other and reanalyzed this subset of genes (n=5,298) with GSEA.

4.2.6 Testing for non-random distribution of p-values

We followed a similar approach to that implemented in Zhernakova *et al*. 2011 (Zhernakova, Stahl et al. 2011) while applying it to genes instead of individual SNPs to test for non-random distribution of association values. For each disease pair we retained all *k* genes that were nominally significant in one disease (p-value < 0.01). We then tested the null hypothesis of a uniform distribution of p-values in the second disease using Fisher's method for combining p-values: $\chi^2_{2k} = -2\sum_{i=1}^{k} \ln(p_i)$, where p_i is the p-value for association of gene *i* in the second disease. Nearby genes in linkage disequilibrium may violate the independency assumption in Fisher's method. We thus performed a separate analysis after removing the latter of the two genes that were within 0.1cM of each other and nominally significant in one disease.

4.2.7 Application of *disPCA* to 31 GWAS datasets

We analyzed a total of 31 GWAS datasets (Helms, Cao et al. 2003; Karamohamed, Golbe et al. 2005; Nichols, Pankratz et al. 2005; Duerr, Taylor et al. 2006; Nair, Stuart et al. 2006; Suarez, Duan et al. 2006; Hunter, Kraft et al. 2007; Matarin, Brown et al. 2007; Saxena, Voight et al. 2007; Scott, Mohlke et al. 2007; Scuteri, Sanna et al. 2007; The Wellcome Trust Case Control Consortium 2007; Boomsma, Willemsen et al. 2008; Cronin, Berger et al. 2008; Harley, Alarcon-Riquelme et al. 2008; Hom, Graham et al. 2008; Li, Wetten et al. 2008; Baranzini, Wang et al. 2009; Barrett, Lee et al. 2009; Nair, Duffin et al. 2009; Sabatti, Service et al. 2009;

Heinzen, Need et al. 2010; Jin, Birlea et al. 2010; Laaksovirta, Peuralinna et al. 2010; Neale, Medland et al. 2010; Remmers, Cosan et al. 2010; Evans, Spencer et al. 2011; Sawcer, Hellenthal et al. 2011; Schizophrenia Psychiatric Genome-Wide Association Study (GWAS) Consortium 2011; Ahn, Ding et al. 2012; Jin, Birlea et al. 2012) that spanned different types of cancers, autoimmune diseases, neurological disorders, psychiatric disorders, type-2 diabetes (T2D), ischemic stroke and body mass index (BMI) (Table 4.S2). Datasets were publicly available, obtained from dbGaP or obtained via collaborations. These datasets had non-overlapping samples and were of European ancestry only. For Wellcome Trust Case Control (WT) related datasets, we distributed controls between the five datasets such that none had overlapping samples. For WT type-1 diabetes, rheumatoid arthritis and Crohn's disease, we obtained further controls from the WT hypertension, cardiovascular disease and bipolar case data (The Wellcome Trust Case Control Consortium 2007). After obtaining gene-level association statistics for 14,018-17,438 autosomal genes for each dataset, we limited our analysis to the 11,927 genes that overlapped all studies. Nineteen of these genes were significantly associated with association method after multiple-testing correction (see above).

4.2.8 Replication of *disPCA*

We tested the replicability of *disPCA* when applied to real GWASs using six datasets for which we had access to the original data (Duerr, Taylor et al. 2006; The Wellcome Trust Case Control Consortium 2007; Baranzini, Wang et al. 2009; Jin, Birlea et al. 2010; Sawcer, Hellenthal et al. 2011; Jin, Birlea et al. 2012). Each dataset was split into independent subsets of equal size (+/- two samples). We then used PLINK's logistic regression (Purcell, Neale et al. 2007) to evaluate association of each SNP to disease risk. We additionally incorporated covariates derived from

EIGENSOFT into the regression analysis (Patterson, Price et al. 2006) to control for population structure. We randomly chose one subset of each of the six datasets for one *disPCA* analysis, and the rest for another. Hence, these two analyses consist of independent samples.

4.3 Results

We first applied *disPCA* to a simulated dataset (Materials and Methods). We varied the number of genes belonging to each category, thereby varying how much power there was to detect pleiotropy between the simulated diseases. *disPCA* was unable to clearly cluster pleiotropic diseases when diseases shared fewer than 40 genes that had p-values below 0.05 (Figure 4.1a-b, 4.S4-4.S6). This can be seen both visually via PCA plots, and via hierarchical clustering based on the Euclidean distance between datasets in the presented space of the first two principal components (PCs) (Figure 4.1, 4.S4-4.S6). When diseases are indeed clustered by their simulated pleiotropy according to *disPCA* (Figure 4.1b), the first two PCs explain a similar fraction of the variance (Figure 4.1c), which may increase or decrease depending on the number of genes contributing to pleiotropy (Figure 4.S7). Genes with p-values < 0.05 (Materials and Methods), which contribute to the simulated pleiotropy between diseases, are also enriched for larger loadings (Figure 4.1d-e).

We next applied *disPCA* to diseases for which we had two datasets. We utilized autoimmune diseases (for which we had the most pairs of datasets) and a pair of schizophrenia datasets (as schizophrenia has a high heritability (Kendler and Diehl 1993)). We observed that datasets of the same diseases were generally clustered together (Figure 4.2-4.3). We additionally observed that Crohn's disease is separated from other autoimmune diseases. This result is consistent with previous reports that inflammatory bowel disorders (IBDs) are distinct from other autoimmune disorders (Sirota, Schaub et al. 2009). As in the simulated scenarios, the variance explained by each PC was similar and suggests that less than a hundred genes contribute to the similarity between each dataset.

To test the replicability of the results, we further divided each of the six datasets, for which we had the raw data, into two subsets consisting of the same or similar samples of cases and controls (Materials and Methods). We then performed two *disPCA* analyses, one on a randomly chosen subset of each of the six datasets and another on the remaining subsets. We found that both independent sets produced the same clustering of diseases (Figure 4.S8-4.S9). Loadings for 50 genes with the largest average loading of PC1 and PC2 in each set were also significantly correlated across the replication sets ($r>0.44$, p-value $< 1.2 \times 10^{-3}$, Table 4.S3).

We applied *disPCA* to a final set of 31 datasets, including autoimmune diseases, cancers, obesity related diseases and traits, psychiatric disorders and neurological disorders. As before, the top PCs explain a similar portion of the variance, with the first two PCs capturing interpretable separation of diseases. PC1 splits systemic lupus erythematosus (SLE), celiac disease and one schizophrenia dataset from all other diseases (Figure 4.4). Alternatively, PC2 splits autoimmune diseases from other diseases, and within autoimmune diseases, inflammatory bowel disorders are clustered together (Figure 4.5). Schizophrenia, major depressive disorder, cancers, T2D and neurological disorders lie on the negative end of PC2, while attention deficit hyperactivity disorder (ADHD) and some autoimmune diseases lie near the origin.

As *disPCA* teases out the important genes of shared and distinct pathogenetics across disease datasets, we next investigated which genes strongly contribute to each PC. The result of applying PCA on a matrix of association values (Materials and Methods) is that each resulting PC is simply a linear combination of genes, whereby each gene is assigned a weight for its contribution

to that PC. We retrieved the genes with the top 50 absolute weights for each of the top two PCs underlying Figure 4.4 and tested their disease enrichment (Materials and Methods). The top genes underlying the first PC were significantly enriched for genes associated with lupus and autoimmune related diseases, while genes underlying the second PC were mostly enriched for association to IBD (Table 4.1). These enrichment results are consistent with the separation of studies across each of these 2 PCs with PC1 separating studies of SLE and other autoimmune diseases, and PC2 separating studies of IBD from other diseases. The results were largely unchanged even after filtering genes that were within 0.1cM of each other (Table 4.1) (Materials and Methods).

Though the results of the disease enrichment analysis support that *disPCA* extracts biologically relevant signals, the arbitrary cutoff of the 50 top genes goes against the potential of PCs being linear combinations of all genes. We thus used GSEA (Subramanian, Tamayo et al. 2005), which supports analyzing a pre-ranked list of all genes, to perform pathway enrichment of each PC. GSEA assesses whether genes belonging to a certain pathway are non-randomly distributed in the list of pre-ranked genes. We ranked all genes by the weight in the PC under study. Results of this pathway analysis revealed enrichment for immune related pathways on the first 2 PCs (Table 4.2) at an FDR of 0.25, as suggested by GSEA (Subramanian, Tamayo et al. 2005) (GSEA manual online), though this entails that 1 in 4 of our results are false positives on average. The top two pathways enriched on PC1 were the antigen processing and presentation and the intestinal immune network IgA production pathways, which are crucial immune-related pathways. In particular, intestinal IgA antibodies may have a role in inflammatory bowel disease (Macpherson, Khoo et al. 1996; Bouvet and Fischetti 1999) and celiac disease (Cunningham-

Rundles 2001). On PC2, the most significant pathway was the NOD-like receptor signaling pathway. NOD-like receptors have been associated to CD, while other immune-related genes likely interacting with NOD2 have been associated to UC (Rubino, Selvanantham et al. 2012). Overall, a majority of the pathways are related to the immune systems. For example, the Fc epsilon RI signaling pathway is related to the antibody IgE, which induces inflammatory response (Pearlman 1999). Other pathways are related to neurons (i.e. the neurotrophin signaling pathway and the Trk-A pathway). In particular, the neurotrophic factor *BDNF* (brain-derived neurotrophic factor), which is a part of the neurotrophin pathway, has been previously associated to Alzheimer's, Parkinson's disease and depression (Momose, Murata et al. 2002; Ventriglia, Bocchio Chiavetto et al. 2002; Sen, Nesse et al. 2003). More recently, an intronic variant in this gene has also been associated to BMI (Berndt, Gustafsson et al. 2013). These associations may explain the separation of neurological, psychiatric and BMI studies on PC2. Because similar genes are sometimes also physically located closer together, we reran GSEA after filtering genes that were within 0.1cM of each other (Materials and Methods). The top two pathways on the first PC remained significant, while only the top pathway in PC2 remained significant (Table 4.S4).

Many autoimmune diseases share associations from the HLA region. We thus reran *disPCA* after removing all genes in and around the HLA region, and found a slightly different visual PCA map (Figure 4.6). SLE and celiac disease were no longer distinguished from other autoimmune diseases and instead lay near the origin. PC1 now differentiated IBD from other diseases, and PC2 distinguished vitiligo from schizophrenia. This was further supported by clustering results on the first two PCs (Figure 4.S10). A GSEA analysis of the PC loadings retained the NOD-like receptor signaling pathway on PC1 instead of PC2 (Table 4.3). Analysis of PC2 loadings

revealed additional immune related pathways that were not enriched in our previous analysis including the HLA region.

Our findings that PC1 splits some autoimmune diseases, and a schizophrenia study from studies of other diseases prompted us to further explore the shared pathogenetics between diseases by testing for the non-random distribution of gene-based p-values in one disease based on their nominal significance in another disease (Materials and Methods). Generally, association statistics are non-randomly distributed when considering most pairs of autoimmune diseases, i.e. testing for non-random distribution in one autoimmune disease dataset based on significance in another autoimmune disease dataset (Figure 4.7). As a control, we tested for non-random distribution for a random set of genes and found that no disease pair was significant for non-random distribution (Figure 4.S11). Our results reported a similar story as observed via *disPCA*. Genes nominally significant in rheumatoid arthritis, type-1 diabetes and ankyolosing spondylitis were non-randomly distributed in SLE and vice versa. We also found that genes nominally significant for one schizophrenia study were non-randomly distributed in a number of autoimmune diseases (Figure 4.7). These signals remained even after genes within 0.1cM of another gene were removed (Figure 4.S12) (Materials and Methods).

4.4 Discussion

In this study we introduced a new method, *disPCA*, to explore the shared pathogenetics of various diseases and disease classes based on GWAS data. PCA has been widely used in population and medical genetics. Applied to genome-wide genotyping data, it can recapitulate European geography (Novembre, Johnson et al. 2008), has been used as a tool to assess and correct for population stratification in GWAS (Patterson, Price et al. 2006; Price, Patterson et al. 2006) and has also been proposed as a tool for reducing the dimensionality of multiple phenotypes for association analysis (Klei, Luca et al. 2008). Our *disPCA* method considers PCA on a different type of matrix, whereby different GWASs are studied in the space of all genes. It can group GWASs of different diseases together based on gene-level association statistics, while accounting for biases due to heterogeneity in sample size, association method, genotyping array and other confounders between studies. This implementation of PCA weighs genes differently on each PC in a manner that distinguishes between diseases. Hence, the higher the level of shared pathogenetics between diseases, the closer they will be in PC space. This is in contrast to a previous method that weights all SNPs equally (Sirota, Schaub et al. 2009). In general, a correlation-based method is less powerful since the correlation between studies across all genes is low, even when the same disease is studied. For example, the correlation coefficient between the –log10 p-values of the two CD studies is 0.048, and it is 0.063 and 0.031 between ulcerative colitis and each of the two CD studies. Furthermore, the highest correlation between pairs of datasets was obtained for schizophrenia (0.13, p-value=2.2×10^{-16}) while the lowest was obtained for type-2 diabetes (0.0031, p-value=0.73). These results show that there is less power when aggregating information across all genes and that *disPCA* is able to tease out and weigh the suitable set of genes underlying shared pathogenetics.

Though *disPCA* is designed to uncover shared disease etiology between diseases, other sources of correlation between datasets can also contribute to *disPCA*. Potential confounders include population structure, shared samples between datasets and technical artifacts. To minimize the impact of these confounders *disPCA* was only applied to studies of individuals with European ancestry and datasets that had no overlapping case or control data. We additionally accounted for technical artifacts introduced by the genotyping array, association method and sample size by regressing out variation in the data attributed to these sources (Materials and Methods). Though we cannot account for other potential confounders that are unknown, the remaining correlation between studies is likely to be due to a shared disease etiology.

We applied *disPCA* to data from 31 GWAS that cover a range of diseases in four main classes: autoimmune diseases, cancers, neurological disorders and psychiatric disorders. We additionally analyzed GWASs on T2D, BMI and ischemic stroke. We first observed that different studies of the same diseases tend to lie closer together on the lead PCs (Figure 4.2). This is in support of studies of the same disease replicating many of the same signals of associations when samples are of similar ancestry. We additionally find that *disPCA* positions diseases within the same class closer together (Figure 4.4). This was especially the case for the major types of IBDs (i.e. Crohn's disease and ulcerative colitis), which clustered close together (Figure 4.5). Between the different disease classes, *disPCA* found overlap between non-autoimmune diseases and traits, and suggests a connection between schizophrenia and some autoimmune diseases.

Using the weightings of genes on each of the leading PCs, we performed disease and pathway

enrichment analysis. We found that PC1, which mainly splits autoimmune disorders from each other, is significantly enriched for genes associated to immune and autoimmune disorders. PC2, which splits IBD studies from studies of other diseases, is significantly enriched for genes in some inflammatory related pathways and genes associated with IBD. Furthermore, some neuron-related pathways were associated to loadings on PC2. In particular, abnormal neurotrophins levels in the brain have been associated to schizophrenia (Durany, Michel et al. 2001; Buckley, Mahadik et al. 2007). Excluding the HLA region revealed significant enrichment for genes in other immune-related pathways. Though the specific analysis presented in this paper focused on the top two PCs, further PCs estimated by *disPCA* can be examined. For example, PC4 of *disPCA* on all GWASs distinguishes rheumatoid arthritis from other diseases (Figure 4.S13). Pathway enrichment analysis highlighted the calcineurin pathway (FDR = 0.182), which involves t-cell activation. Additionally, though schizophrenia and vitiligo datasets are further apart on the first two PCs, each pair of datasets is clustered closer together on PC3 and PC4. Altogether these results support the validity of the enrichment analysis based on *disPCA*. The analysis in turn also raises new hypotheses of disease etiology by pointing to additional pathways and enrichment for other diseases that were not previously observed.

Prompted by the results of *disPCA*, we further explored shared pathogenetics by testing for the non-random distribution of association statistics between pairs of disease studies (Figure 4.7). Autoimmune diseases show non-random distribution of association statistics with one another. Interestingly, genes nominally associated with one of the schizophrenia studies were non-randomly distributed in studies of several autoimmune diseases (i.e. ankyolosing spondylitis, systemic lupus erythematosus, and T1D). This supports our *disPCA* results above and is in

agreement with epidemiological evidence for a relationship between autoimmune diseases and schizophrenia (Benros, Eaton et al. 2013). This relationship was not observed in the other schizophrenia study, which may be due to a number of factors such as a lack of power. Though, if indeed autoimmune diseases and schizophrenia share disease etiology, then just as one would not include individuals with ulcerative colitis as controls for a Crohn's disease GWAS since they both are IBDs, one should also be wary of including individuals with schizophrenia as controls in an autoimmune GWAS (and vice versa) as doing so may decrease power in loci implicated in both diseases.

Finally, we make a few recommendations for future applications of *disPCA* to additional studies: (1) Biases can be introduced when studies share sample data; (2) As *disPCA* maximizes variance across diseases, genes that are implicated in all analyzed diseases will not contribute to the lead PC as they do not distinguish diseases from each other; (3) While here we only focused on using the strength of association and on gene-level signals, the method itself is highly flexible. One can further utilize the direction of association (protective versus deleterious), the heritability at each locus (Gusev, Bhatia et al. 2013), an analysis at the pathway-level or in linkage-disequilibrium blocks, and/or include other non-genic functional elements; (4) *disPCA* can be used to generate new hypotheses, which can then be tested by conducting more focused association studies in independent data or by using its output to better combine different diseases in an independent meta-analysis. In conclusion, *disPCA* offers users a unique general overview of the disease landscape by studying their distinct and shared pathogenetics and flagging pathways and genes for further investigation. *disPCA*'s flexibility and computational efficiency proves itself as an excellent tool to be applied to additional diseases and disease classes to further our knowledge of

shared pathogenetics.

4.5 Figures

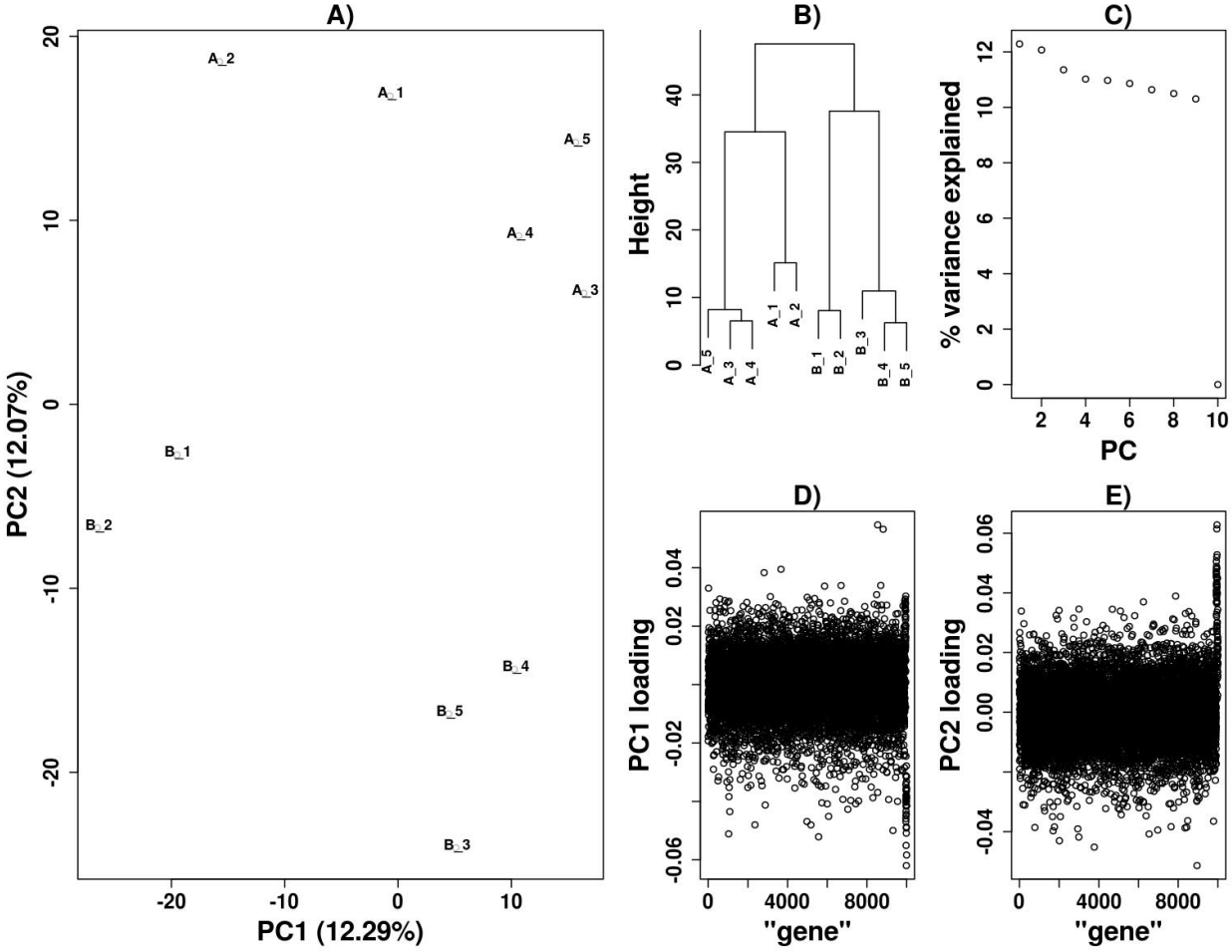

Figure 4.1. disPCA of ten simulated diseases.
The p-values for ten diseases were simulated for 10,000 genes (Materials and Methods). Class *A* diseases had p-values uniformly distributed between 0 and 0.05 for 40 genes, while two diseases from class *A* (*A_1, A_2*) and two diseases from class *B* (*B_1, B_2*) had p-values similarly distributed for a separate 40 genes (Materials and Methods). All other diseases had p-values that were randomly distributed between 0 and 1. A) The simulated data is displayed on PC1 and PC2. PC1 separates (*A_1, A_2, B_1, B_2*) from all other diseases, while PC2 separates class *A* diseases from class *B* diseases. B) Dendrogram derived from a clustering analysis based on the Euclidean distance between datasets in the space of the first two PCs (represented as the height of the branches). C) PC1 and PC2 account for a similar amount of variance. D) Loadings for each gene are displayed sequentially for PC1. The 40 genes contributing to pleiotropy between the two diseases in each class are enriched for larger absolute loadings. E) Similar to (D), with loadings for PC2 displayed. The 40 genes contributing to correlation between diseases in each class and are also enriched for larger loadings.

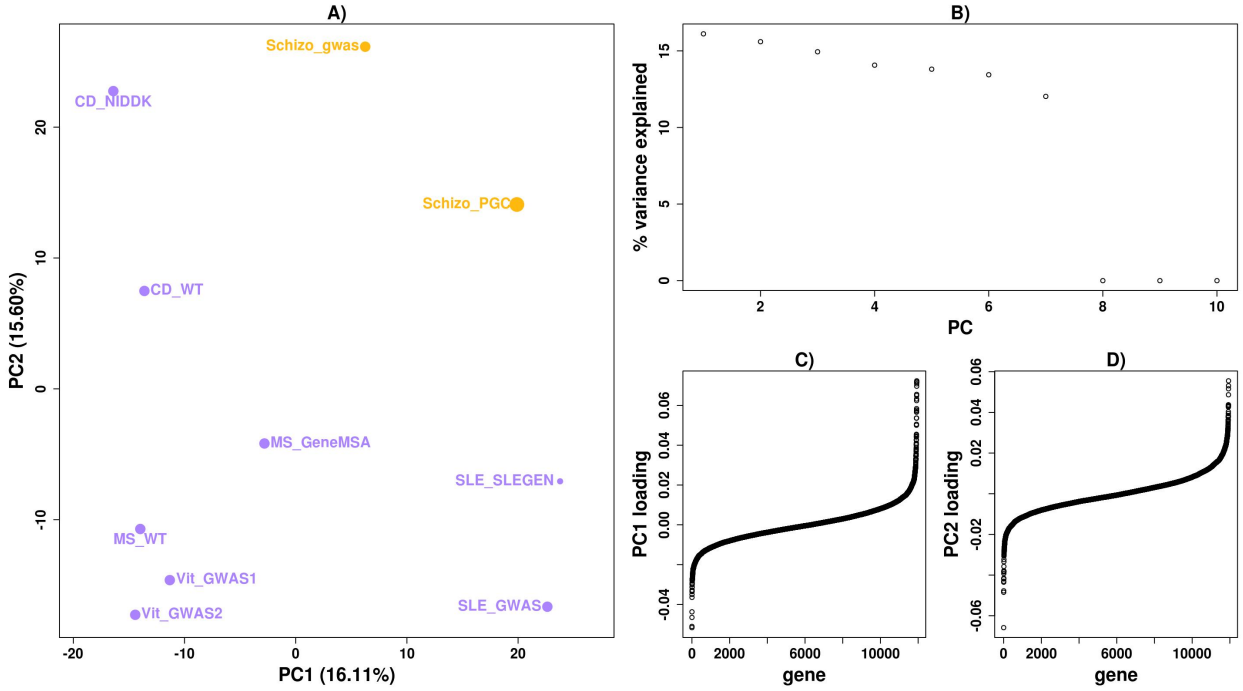

Figure 4.2. disPCA of datasets of the same disease.
A) Pairs of datasets of the same autoimmune diseases and schizophrenia are displayed on PC1 and PC2. Dataset labels are indicated in the form of *disease-type_ study-name*. The size of points is proportional to the sample size of the original study (Table 4.S2). Diseases include systemic lupus erythematosus (SLE), vitiligo (Vit), multiple sclerosis (MS), schizophrenia (Schizo) and Crohn's disease (CD). Datasets of the same diseases tend to lie closer together on PC1 and PC2. B) The portion of variance explained by each PC is displayed. Three additional PCs explain 0% of the variance corresponding to the number of confounders we accounted for (Materials and Methods). C) The weightings for genes on PC1 are displayed and ordered according to their weights. D) Similar to (C) where loadings are for PC2.

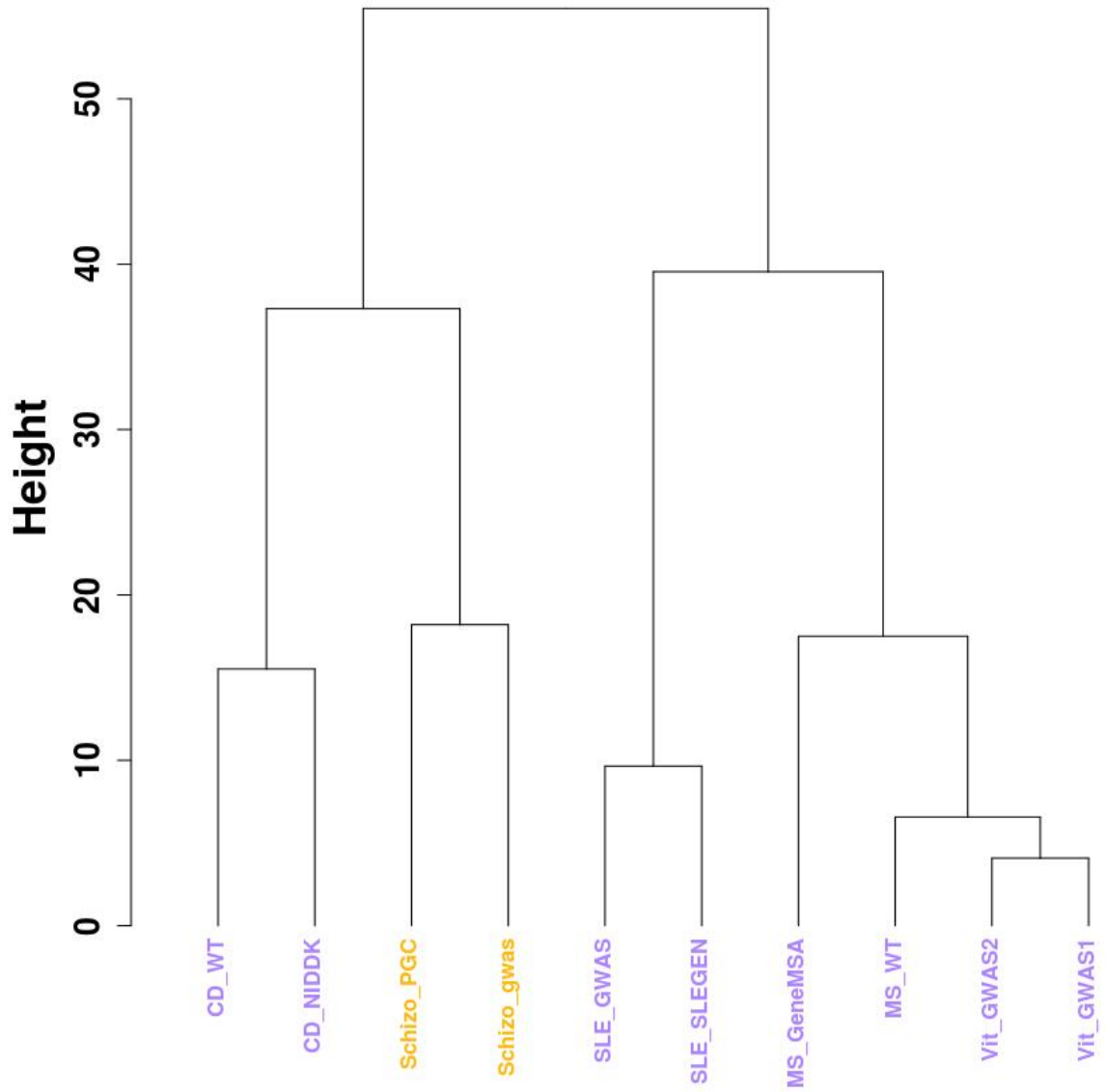

Figure 4.3. Dendrogram of datasets of the same disease.
Hierarchical clustering was applied to the Euclidean distance between datasets in the first two PCs presented in Figure 4.2 (Materials and Methods). The height of the branches represents the Euclidean distance between datasets in the space of the first two PCs. Datasets of the same diseases are clustered together.

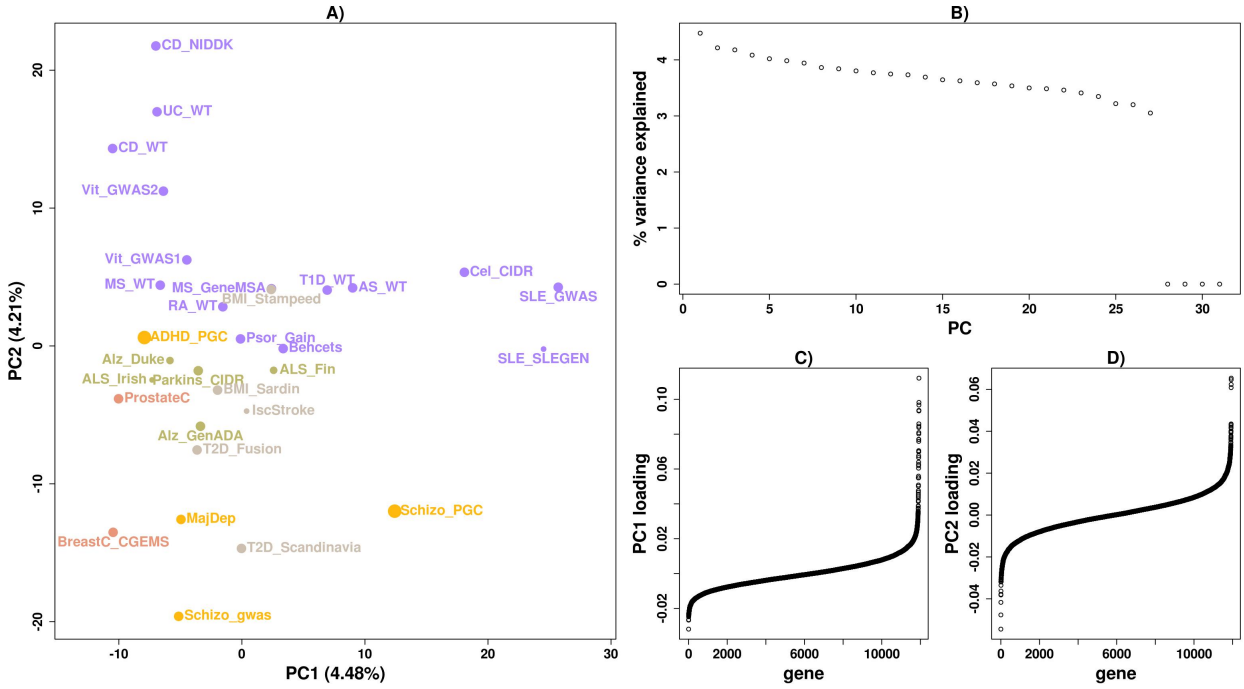

Figure 4.4. disPCA of all diseases and traits.
A) Autoimmune diseases (purple), cancers (pink), psychiatric disorders (yellow), neurological disorders (green), and other diseases and traits (grey) are shown on PC1 and PC2. PC1 accounts for 4.48% of the variance, while PC2 accounts for 4.21%. Additional diseases include Alzheimer's disease (Alz), amyotrophic lateral sclerosis (ALS), ankylosing spondylitis (AS), attention deficit hyperactivity disorder (ADHD), Behcet's disease (Behcets), body mass index (BMI), breast cancer (BreastC), celiac disease (CeliacD), ischemic stroke (IscStroke), major depression (MajDep), Parkinson's disease (Parkin), prostate cancer (ProstateC), psoriasis (Psor), rheumatoid arthritis (RA), type-1 diabetes (T1D), type-2 diabetes (T2D), ulcerative colitis (UC). PC1 clusters celiac disease and SLE together, while PC2 separates inflammatory bowel diseases from other diseases and traits. B) The portion of variance explained by each PC is displayed. Three additional PCs explain 0% of the variance corresponding to the number of confounders we accounted for (Materials and Methods). C) The weightings for genes on PC1 are displayed and ordered according to their weights. D) Similar to (C) where loadings are for PC2.

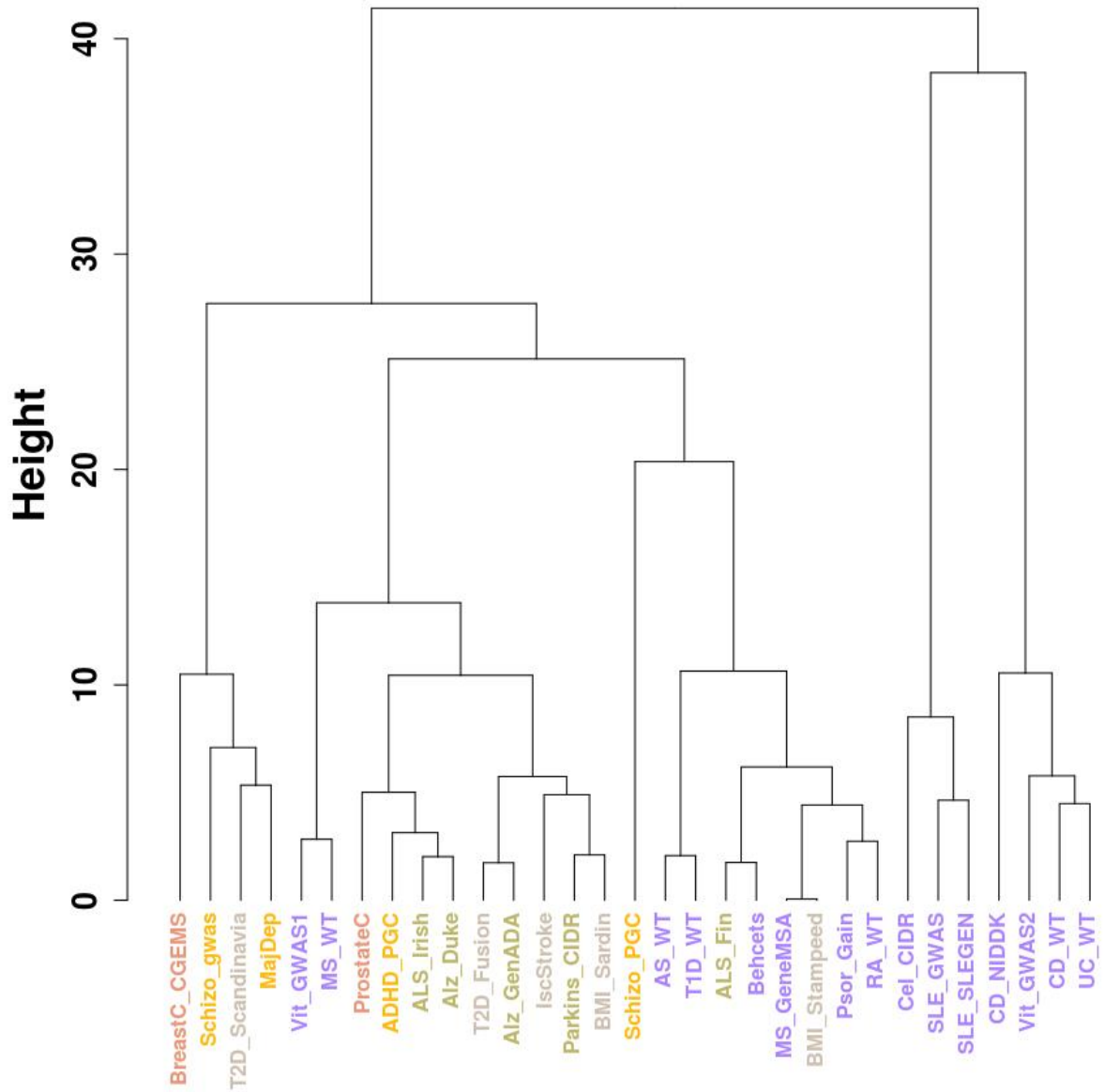

Figure 4.5. Dendrogram of datasets of all diseases and traits.
Dendrogram derived from hierarchical clustering analysis applied to distance (in PC space) between datasets presented in Figure 4.4. Inflammatory bowel diseases are clustered together, in addition to SLE and celiac disease.

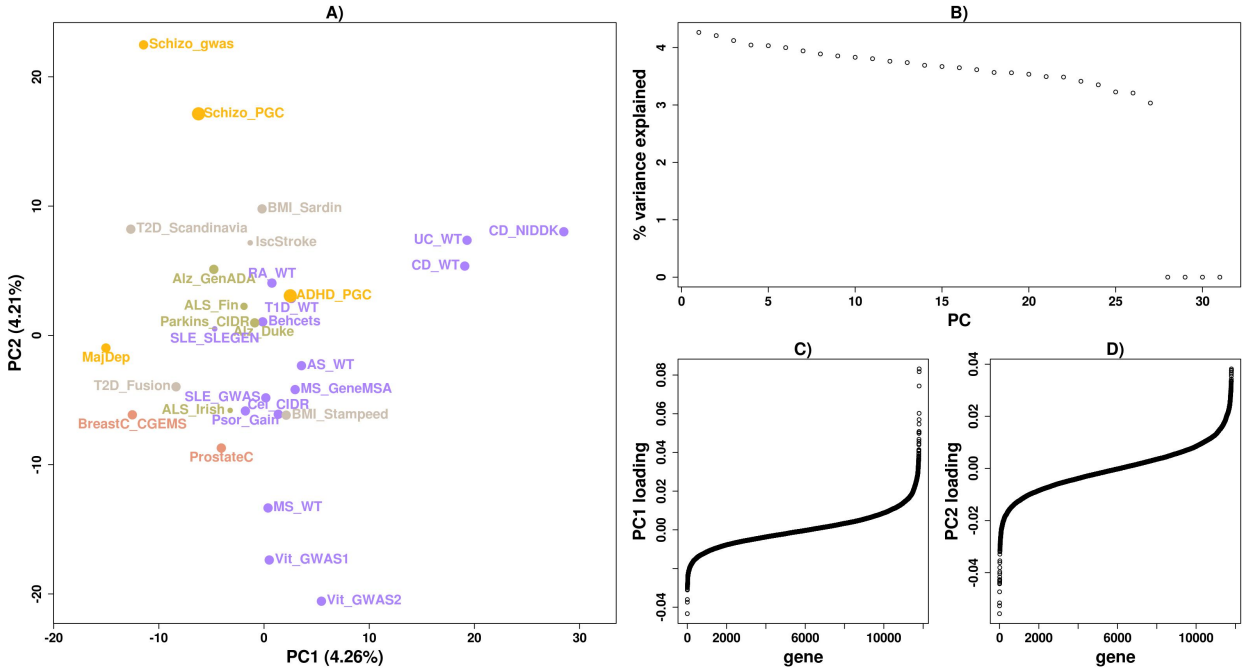

Figure 4.6. disPCA of all diseases and traits excluding the HLA and surrounding region. A) Similar to Figure 4.4 where genes in the HLA and surrounding region (Materials and Methods) were removed. Though IBD remains separated as in the original *disPCA*, the clustering of T1D and SLE is no longer captured by the top two PC's. B) The portion of variance explained by each PC is displayed. C) The weightings for genes on PC1 are displayed and ordered according to their weights. D) Similar to (C) where loadings are for PC2.

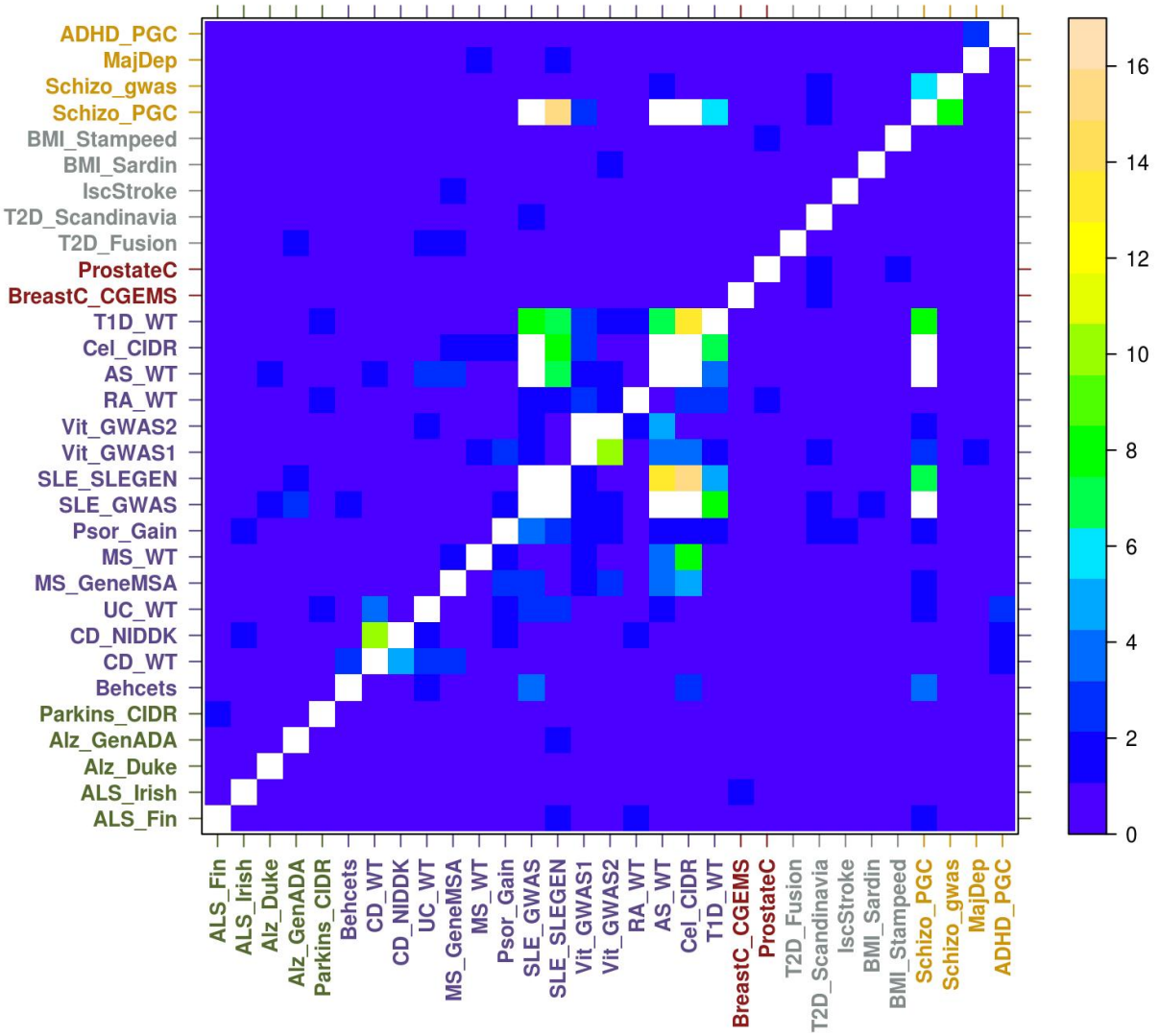

Figure 4.7. Non-random distribution of genes for all analyzed datasets from Figure 4. Genes nominally significant for diseases on the *y*-axis were tested for non-random distribution in diseases on the *x*-axis (Materials and Methods), with $-\log_{10}$ presented on the color scale on the right. White entries denote p-values < 1×10^{-17}. The most significant results are for pairs of similar diseases and between pairs of autoimmune diseases. In addition, pairs between some autoimmune diseases and schizophrenia also display significant results.

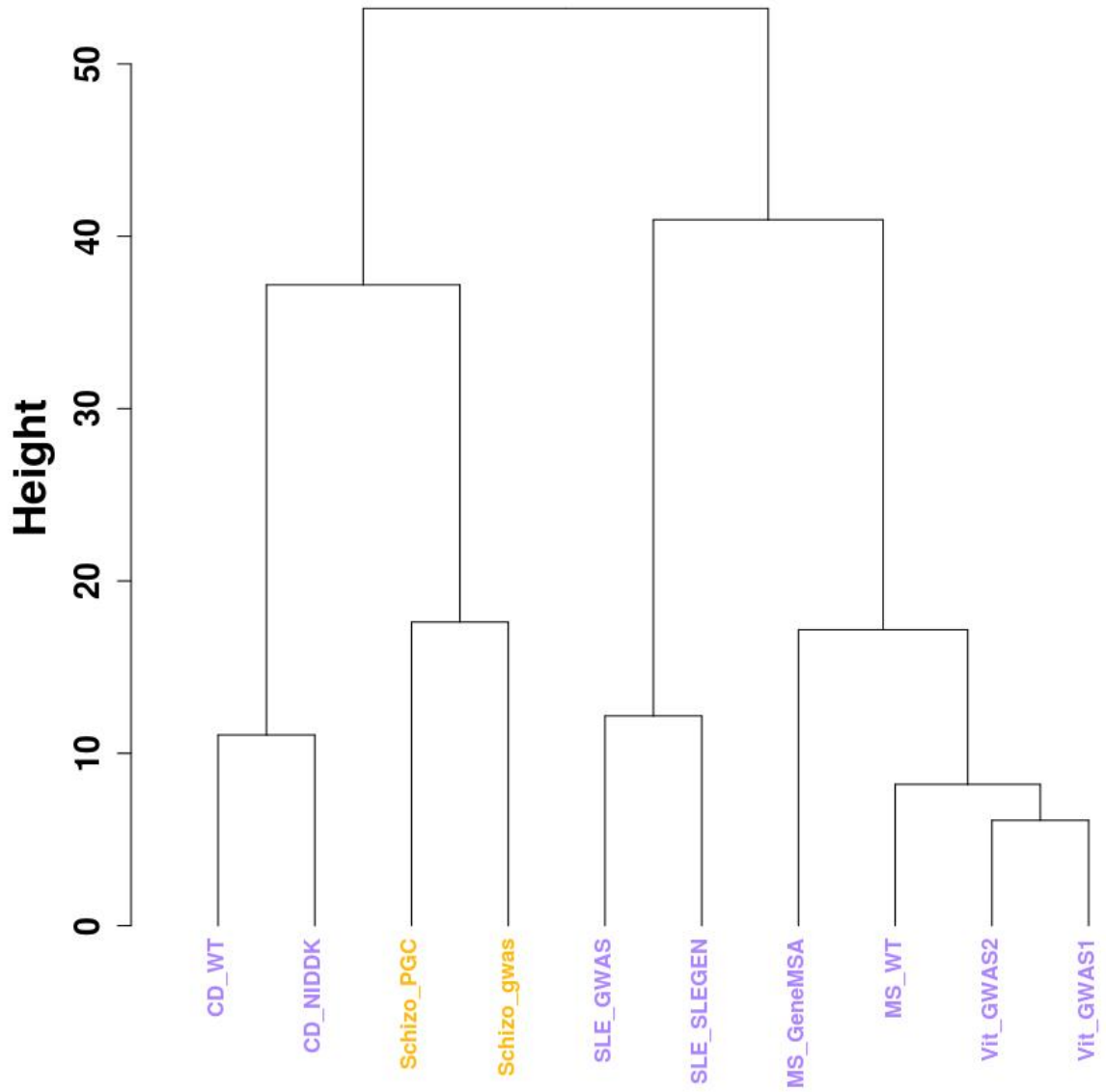

Figure 4.S1. Dendrogram derived from clustering analysis of datasets of the same diseases using physical distance mapping.
SNPs were mapped to genes if they were within 10kb of the gene. Clustering analysis of resulting *disPCA* revealed the same clusters as *disPCA* with genetic coordinates (Figure 4.3).

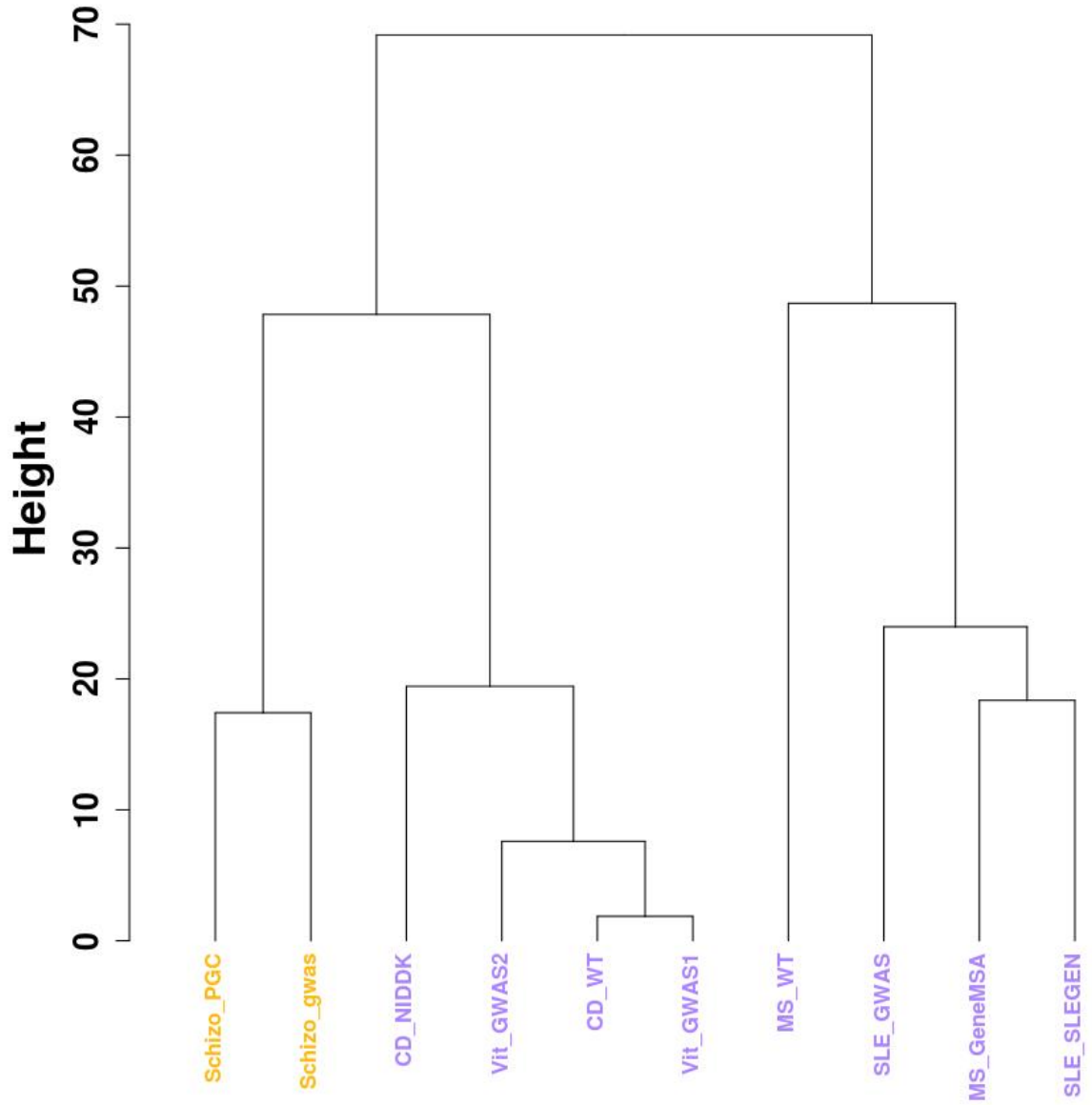

Figure 4.S2. Dendrogram of clustering analysis of datasets of the same diseases with the truncated product method.
Similar to Figure 4.3, with the truncated product method used to combine SNP p-values per gene.

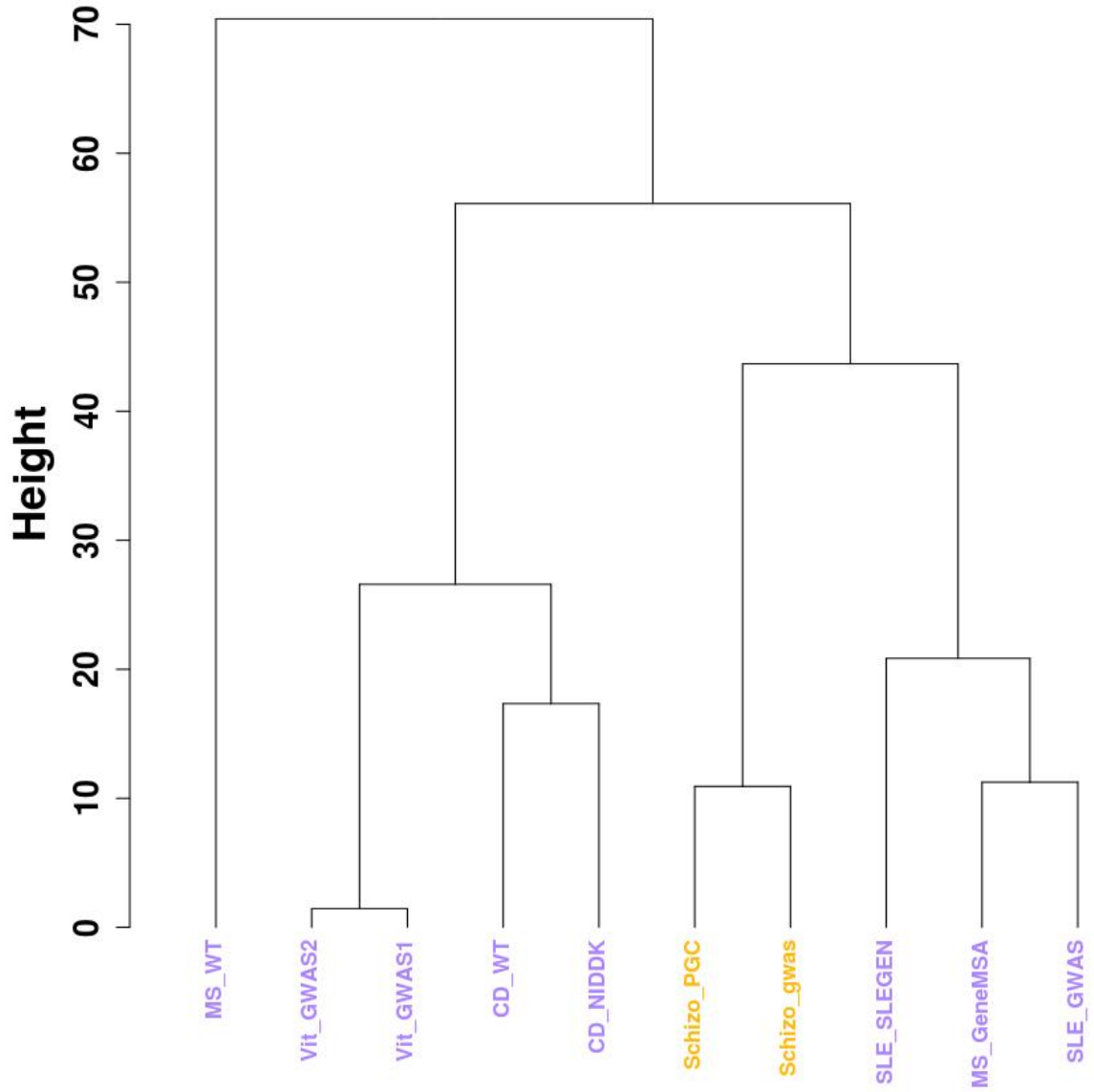

Figure 4.S3. Dendrogram of clustering analysis of datasets of the same diseases with truncated tail strength method.
Similar to Figure 4.3, with the truncated tail strength method used to combine SNP p-values per gene.

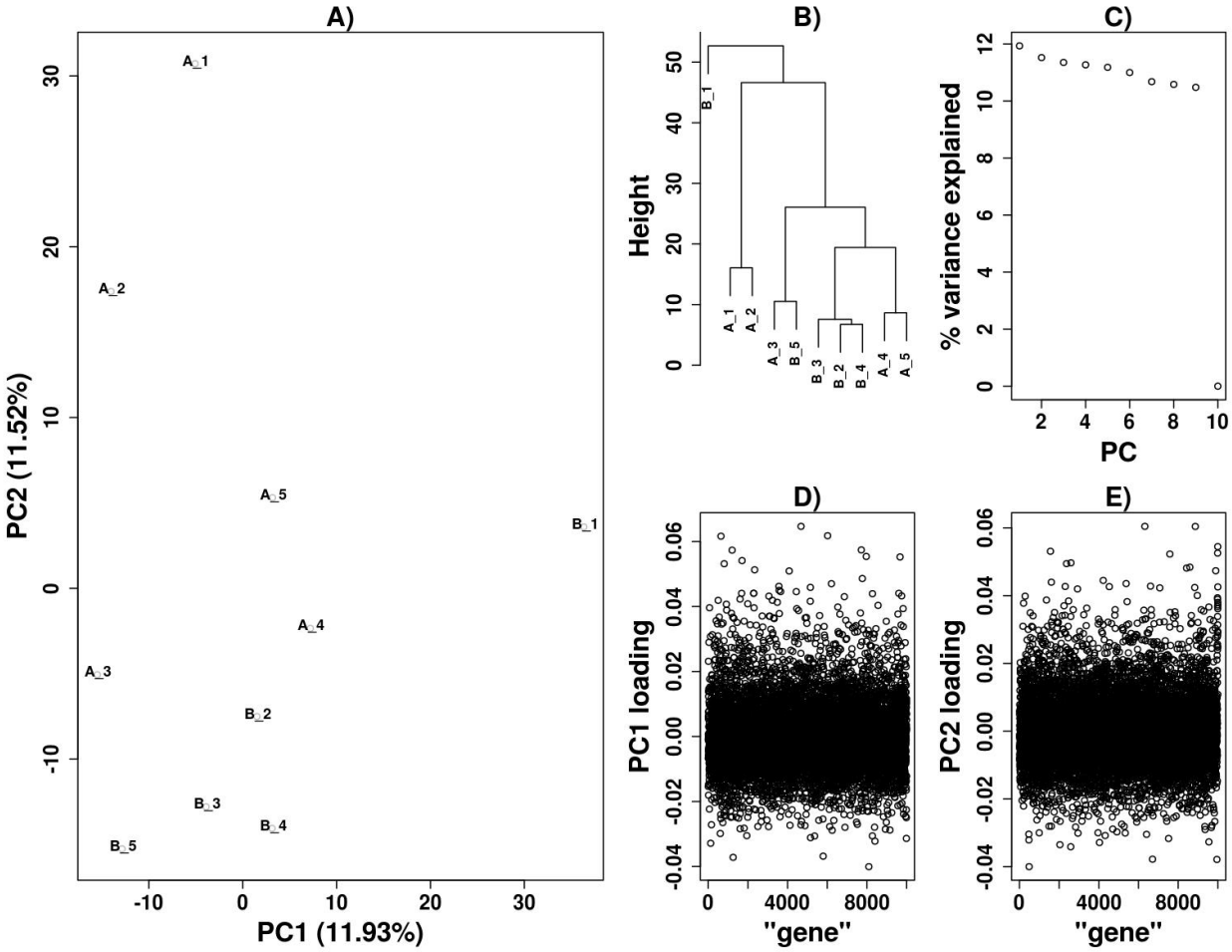

Figure 4.S4. Simulated diseases with ten nominally significant genes.
A) Similar to Figure 4.1 in main text with only ten nominally significant genes for each set of pleiotropic diseases (Materials and Methods). Clustering of the diseases sets is not observed. B) Dendrogram derived from clustering analysis as similarly presented in Figure 4.1b. C) The portion of variance explained by each PC is displayed. D-E) The loadings for PC1 and PC2 are displayed after sorting genes according to their loadings.

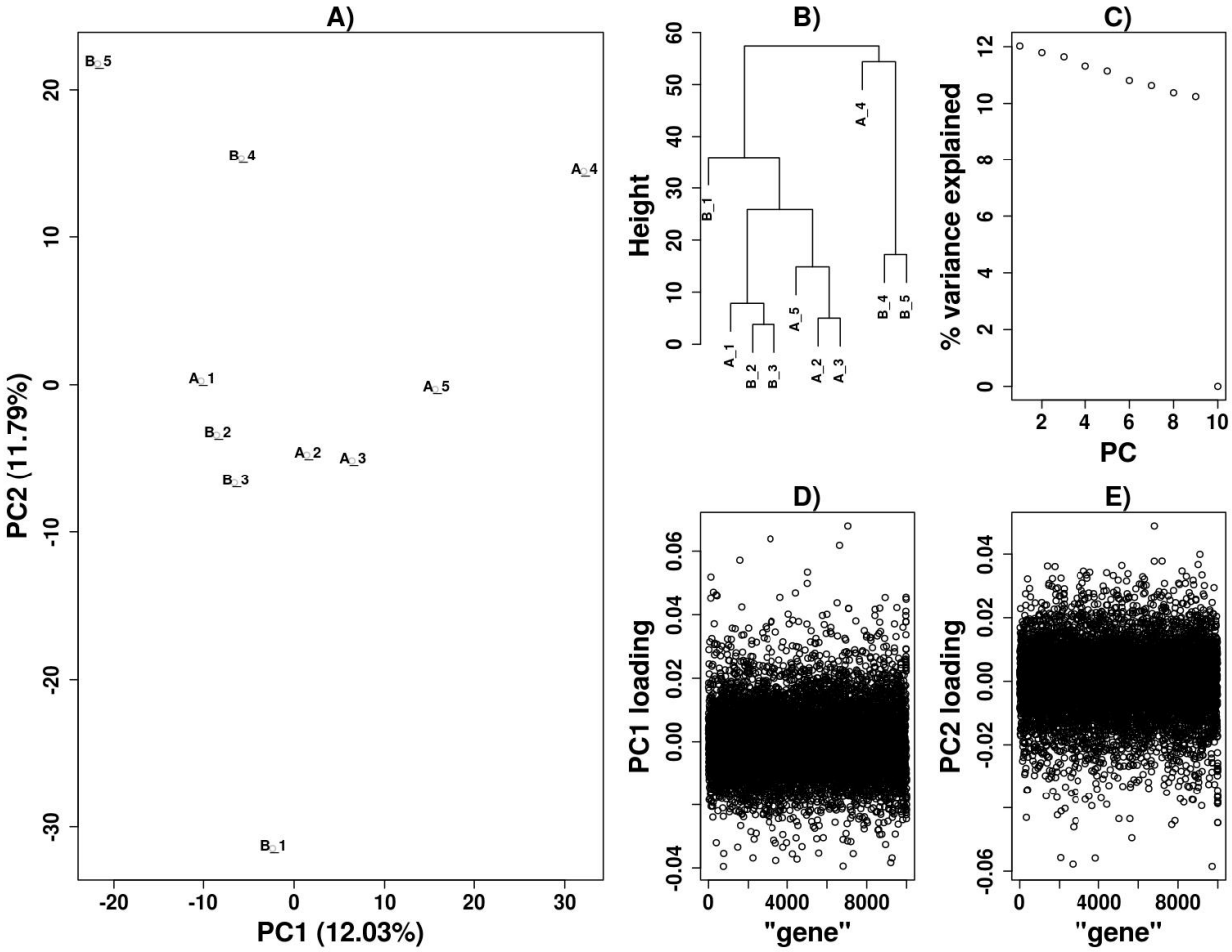

Figure 4.S5. Simulated diseases with twenty nominally significant genes.
A) Similar to Figure 4.1 with twenty nominally significant genes for each set of pleiotropic diseases. As in Figure 4.S2, diseases are not clustering according to the sets though nominally significant genes are enriched for larger absolute loadings (Materials and Methods). B) Dendrogram derived from clustering analysis as similarly presented in Figure 4.1b. C) The portion of variance explained by each PC is displayed. D-E) The loadings for PC1 and PC2 are displayed after sorting genes according to their loadings.

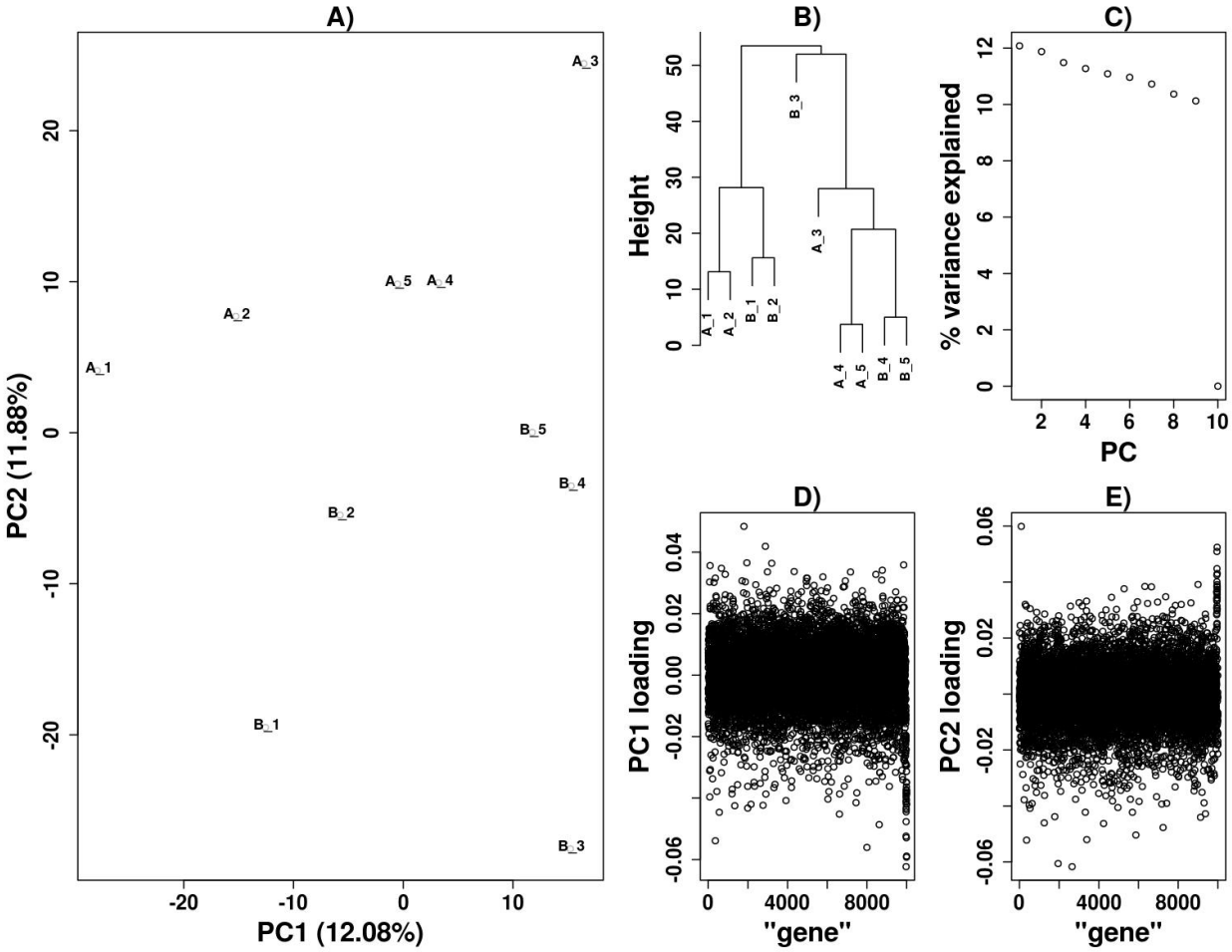

Figure 4.S6. Simulated diseases with thirty nominally significant genes.
A) Similar to Figure 4.1 with thirty nominally significant genes for each set of pleiotropic diseases. The proper clustering of diseases is beginning to emerge. B) Dendrogram derived clustering analysis as similarly presented in Figure 4.1b. C) The portion of variance explained by each PC is displayed. D-E) The loadings for PC1 and PC2 are displayed after sorting genes according to their loadings. Genes with nominally significant p-values are enriched for larger absolute loadings.

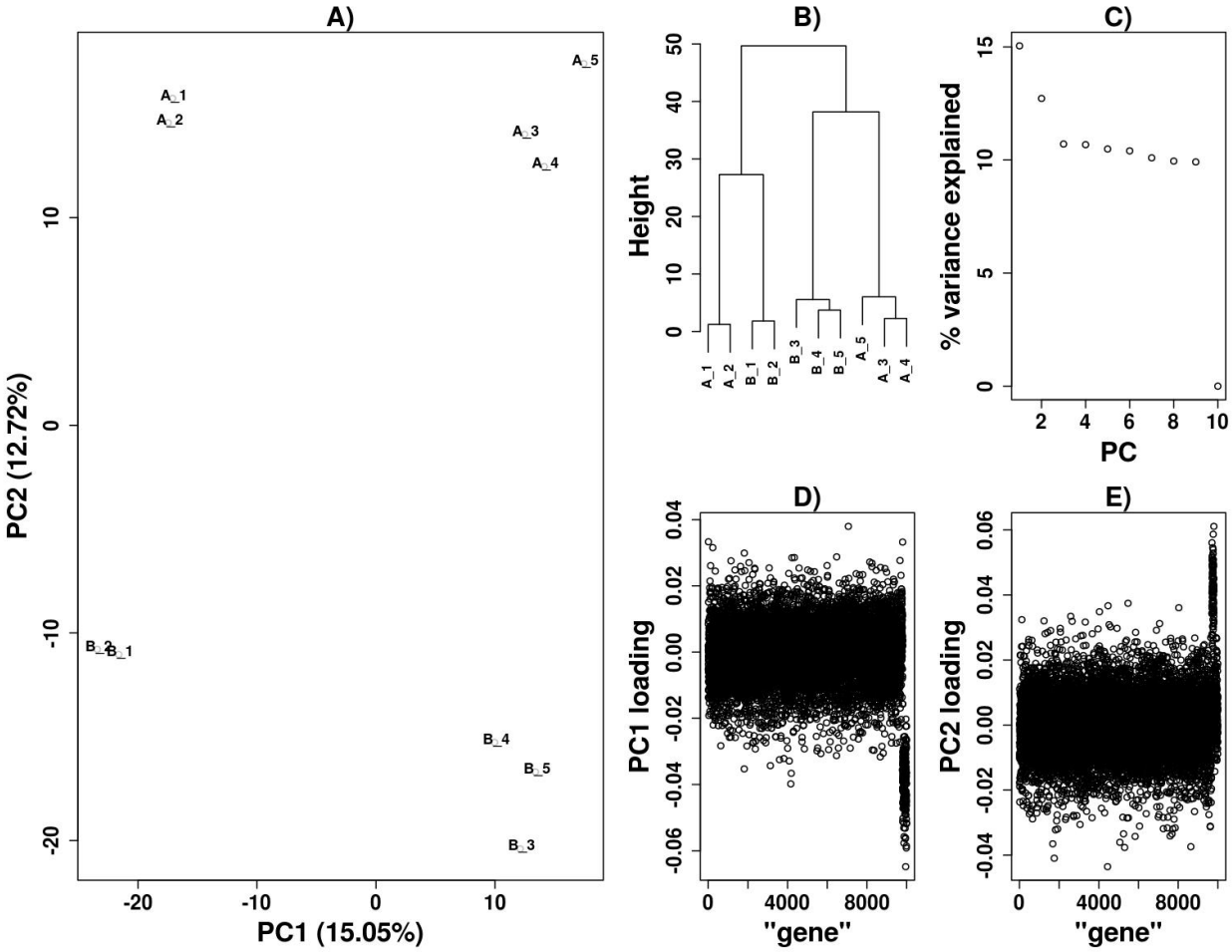

Figure 4.S7. Simulated diseases with 100 and 200 nominally significant genes.
A) Similar to Figure 4.1 with 100 and 200 nominally significant genes for the two sets of pleiotropic diseases. Disease sets are tightly clustered and the first two PCs explain a larger portion of the variance compared to other PCs. B) Dendrogram derived from clustering analysis as similarly presented in Figure 4.1b. C) The portion of variance explained by each PC is displayed. D-E) The loadings for PC1 and PC2 are displayed after sorting genes according to their loadings.

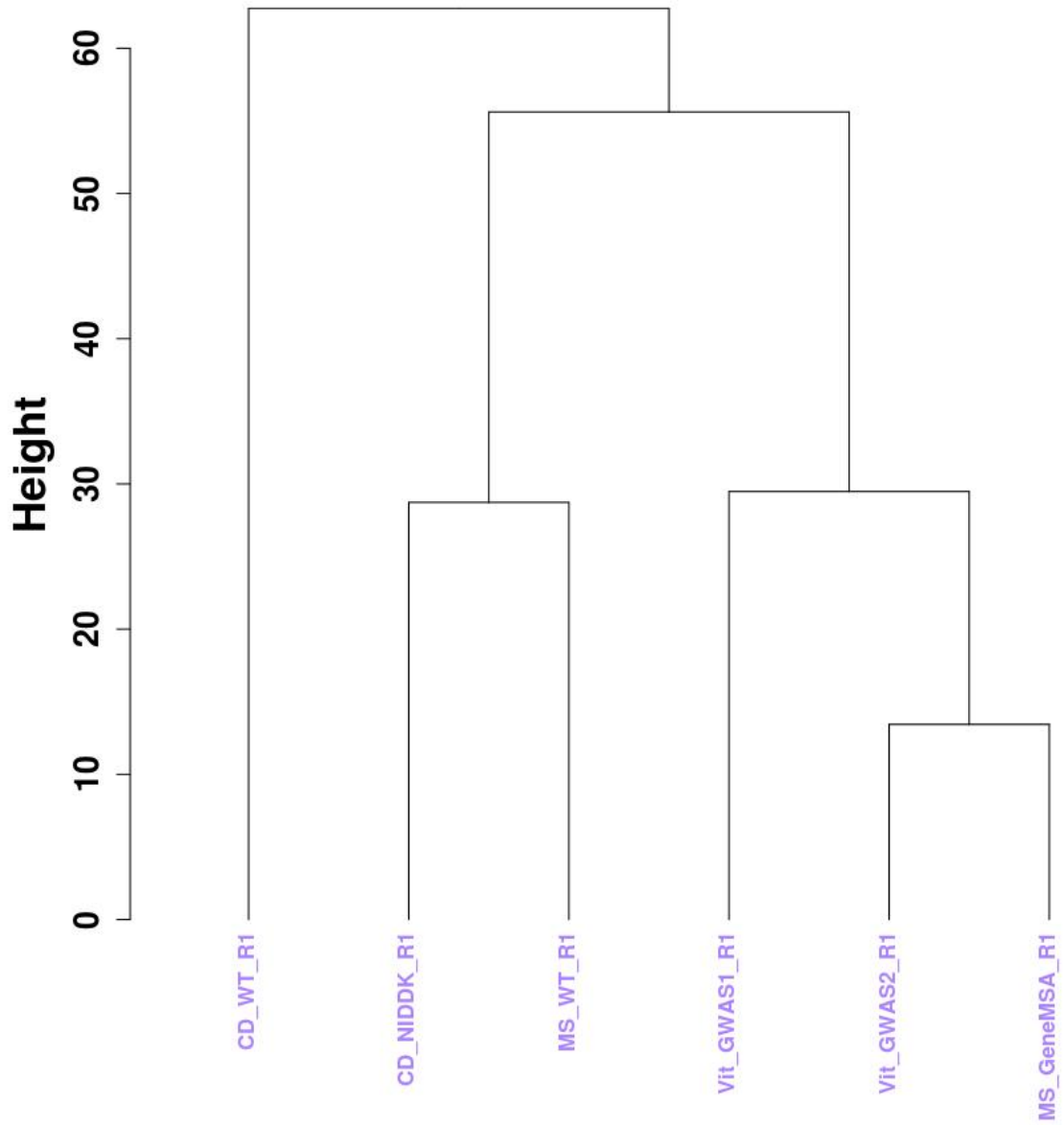

Figure 4.S8. Dendrogram of clustering analysis of Replication Set 1 datasets.
Clustering of the distance in PC space between datasets in Replication Set 1. Diseases include vitiligo (Vit), multiple sclerosis (MS), schizophrenia (Schizo) and Crohn's disease (CD).

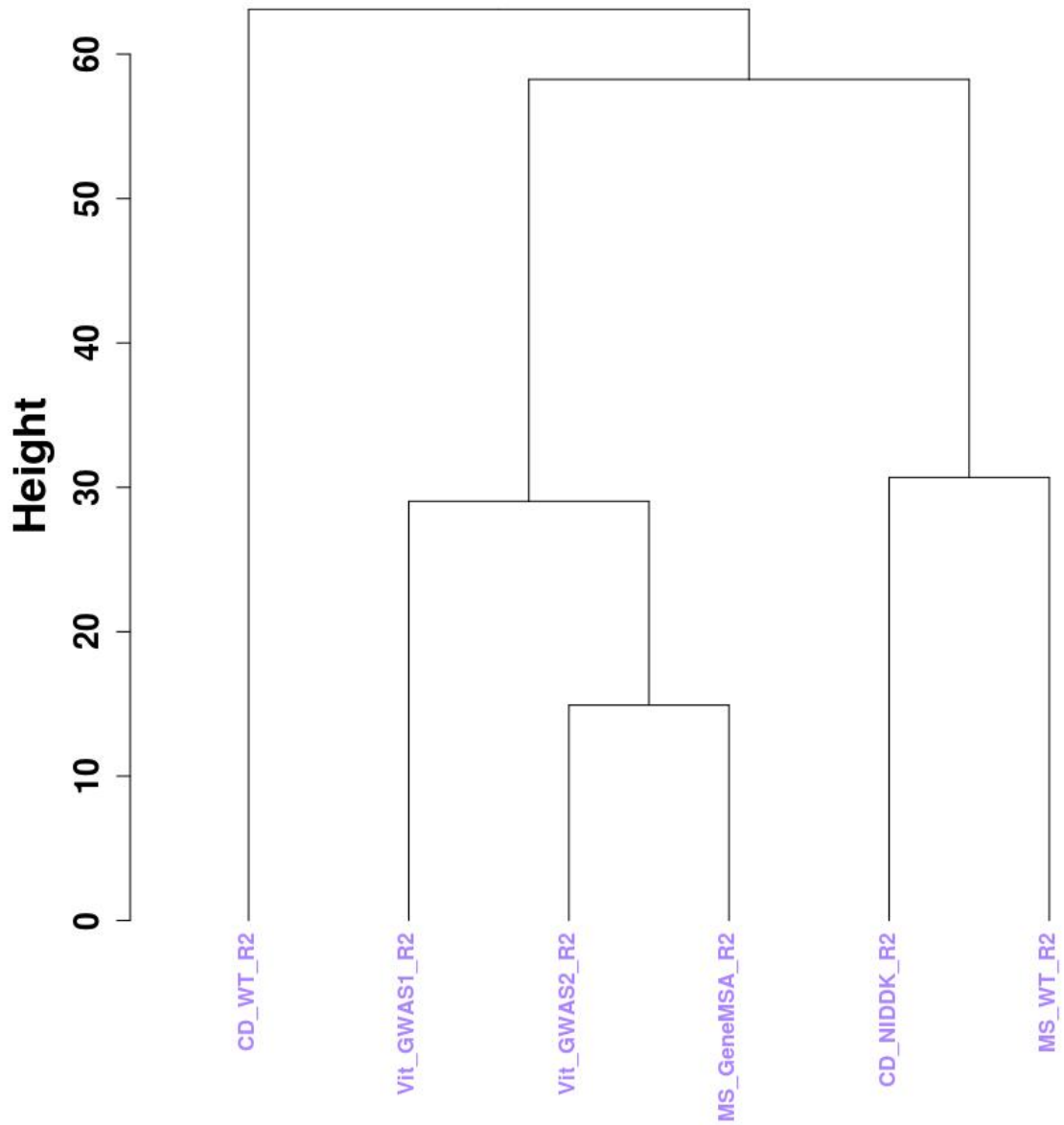

Figure 4.S9. Dendrogram of clustering analysis of Replication Set 2 datasets.
Similar to Figure 4.S8 with datasets from Replication Set 2.

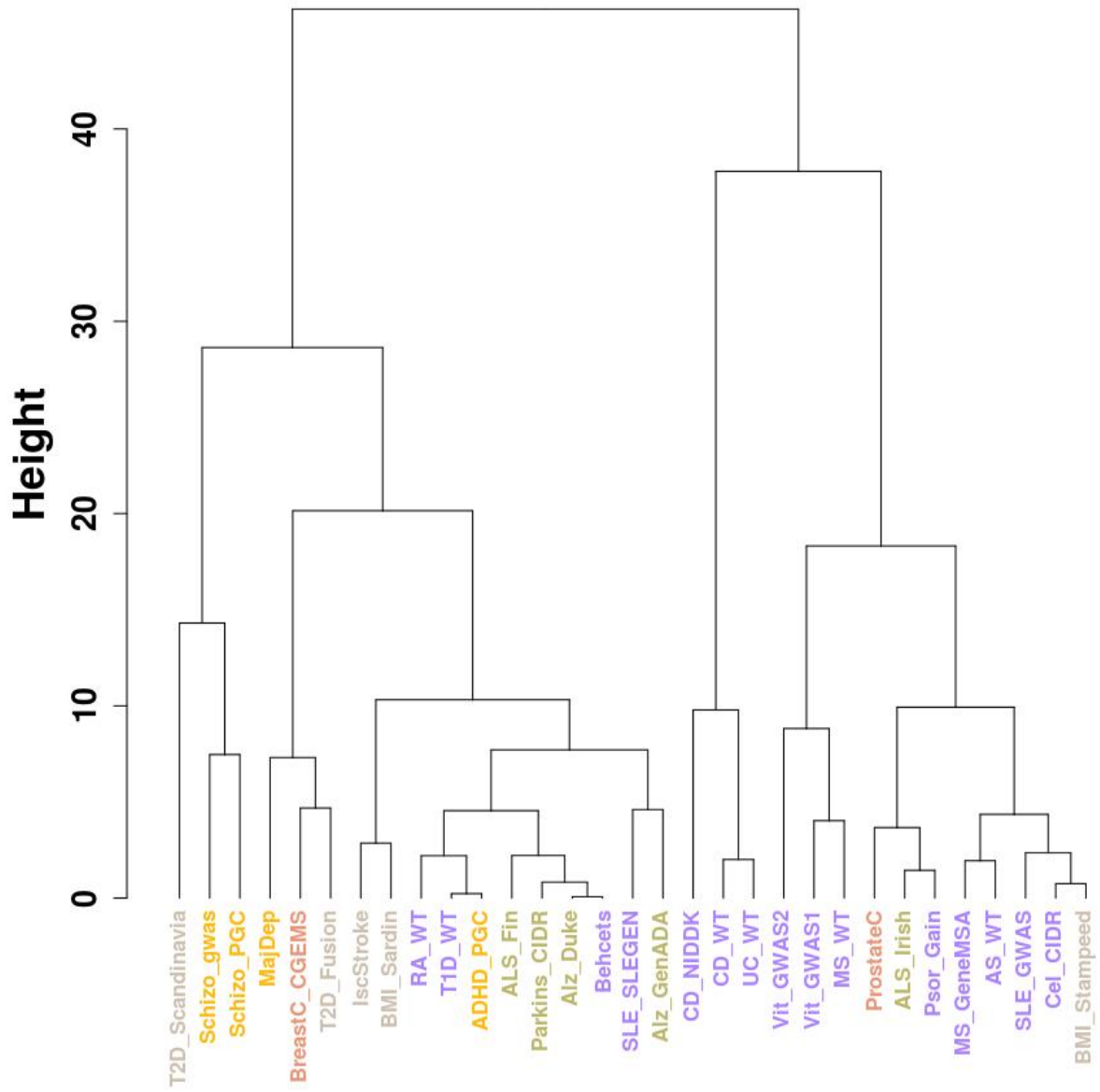

Figure 4.S10. Dendrogram of clustering analysis of all diseases and traits excluding the HLA and surrounding regions.
Figure is similar to Figure 4.5, with clustering analysis of distance between datasets based on the *disPCA* between all diseases and traits presented in Table 4.S2 after removing the HLA and surrounding regions.

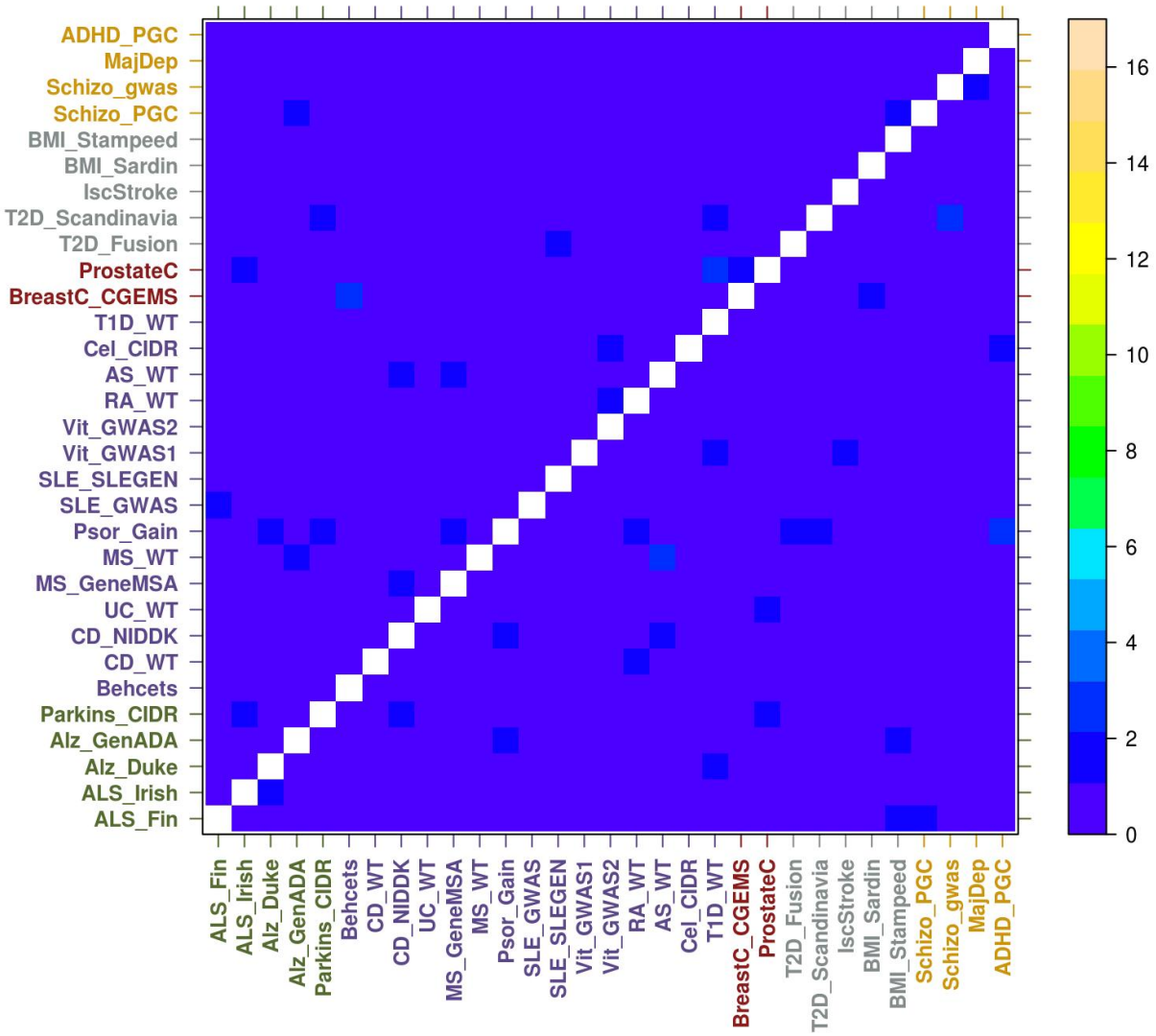

Figure 4.S11. Non-random distribution of randomly chosen genes.
A random subset of genes were chosen to be tested for non-random distribution in diseases on the x-axis, with $-\log_{10}$ presented on the color scale on the right. White entries denote p-values < 1×10^{-17}.

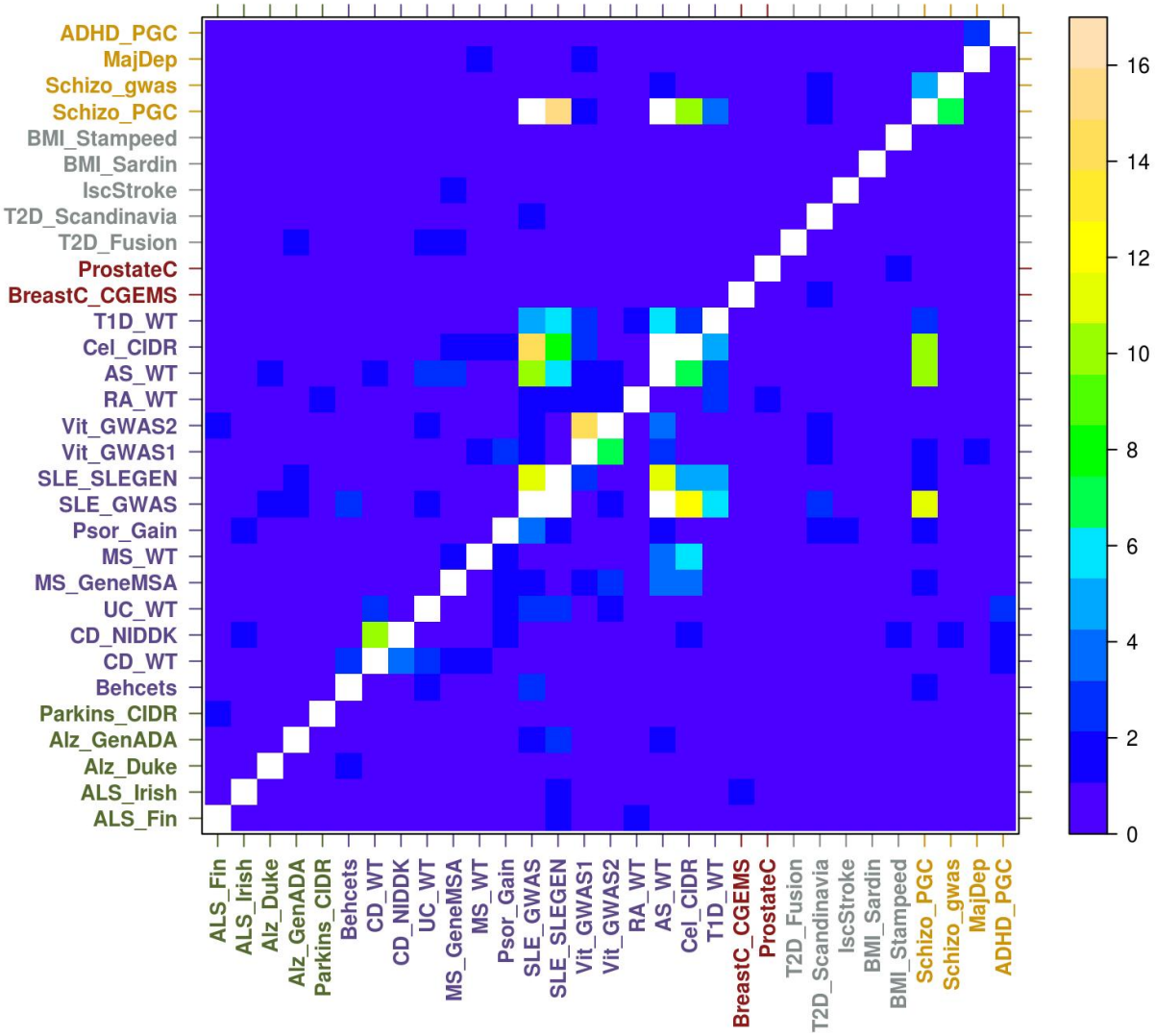

Figure 4.S12. Non-random distribution for distance pruned set of genes.
Genes were filtered such that no two genes were within 0.1cM of another. The remaining subset of genes was then tested for non-random distribution in diseases on the x-axis. The $-\log_{10}$ of the p-value is presented on the color scale and white entries denote p-values $< 1x10^{-17}$. Results are largely similar to the original without filtering of nearby genes.

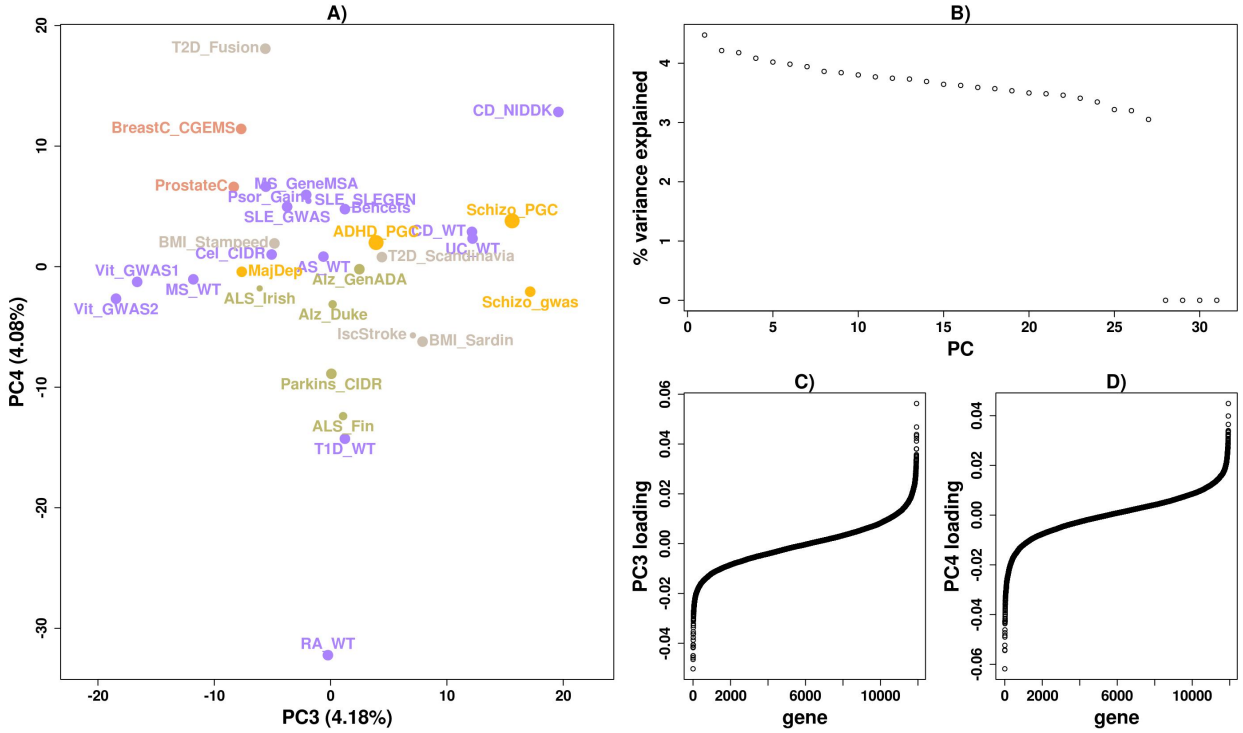

Figure 4.S13. PC3 and PC4 of all diseases disPCA.
Similar to Figure 4.4 with data being presented for PC3 and PC4. A) PC1 accounts for 4.18% of the variance, while PC2 accounts for 4.08%. PC1 clusters celiac disease and SLE together, while PC2 separates inflammatory bowel diseases from other diseases and traits. B) The portion of variance explained by each PC is displayed. C) The weightings for genes on PC1 are displayed and ordered according to their weights. D) Similar to (C) where loadings are for PC2.

4.6 Tables

PC	Disease	P-value*	P-value (distance pruned)*
1	Lupus erythematosus	1.59×10^{-6}	3.0×10^{-8}
	Arthritis	1.72×10^{-6}	>0.01
	Connective tissue diseases	5.00×10^{-4}	>0.01
	Autoimmune diseases	2.6×10^{-3}	2.05×10^{-6}
	Rheumatic Diseases	2.6×10^{-3}	>0.01
	Immune system diseases	6.5×10^{-3}	2.2×10^{-5}
2	Gastroenteritis	5.79×10^{-13}	2.92×10^{-9}
	Crohn's Disease	2.12×10^{-12}	1.73×10^{-8}
	Inflammatory bowel diseases	1.65×10^{-11}	7.53×10^{-8}
	Fistula	4.00×10^{-9}	1.37×10^{-7}
	Gastrointestinal diseases	3.49×10^{-8}	7.16×10^{-8}
	Celiac disease	2.75×10^{-5}	7.8×10^{-6}
	Multiple sclerosis	2×10^{-3}	7×10^{-4}
	Skin diseases, genetic	2.3×10^{-3}	8.1×10^{-3}
	Rheumatic diseases	6.4×10^{-3}	2.3×10^{-3}
	Autoimune diseases	9.6×10^{-3}	2.7×10^{-3}

* Bonferroni adjusted for multiple testing

Table 4.1. Disease enrichment analysis for disPCA (Figure 4.1).
Table shows disease enrichment results for all diseases significantly enriched with an adjusted p-value < 0.01. The distance pruned p-values refers to disease enrichment after removing the latter out of a pair of genes that were within 0.1cM of each other.

PC	Pathway	FDR (q-value)
1	Antigen processing and presentation	0.034
	Intestinal immune network for IgA production	0.042
	Trk-A pathway	0.169
	CK1 pathway	0.213
	DREAM pathway	0.228
	Valine leucine and isoleucine biosynthesis	0.228
	O-glycan biosynthesis	0.243
	Folate biosynthesis	0.246
2	NOD-like receptor signaling pathway	$<1x10^{-4}$
	Intestinal immune network for IgA production	0.074
	Neurotrophin signaling pathway	0.165
	Chemokine signaling pathway	0.195
	Fc epsilon RI signaling pathway	0.232
	Terpenoid backbone biosynthesis	0.232
	JAK-STAT signaling pathway	0.238

Table 4.2. Gene enrichment analysis for disPCA.
Table shows pathways that are enriched in the *disPCA* analysis based on the GSEA analysis.

PC	Pathway	FDR q-value
1	NOD-like receptor signaling pathway	0.006
	Local acute inflammatory response pathway	0.143
2	Proteasome pathway	0.077
	Th1-Th2 pathway	0.102
	Proximal tubule bicarbonate reclamation	0.135
	Adherens junction	0.142
	RNA polymerase	0.171
	CTLA-4 pathway	0.173

Table 4.3. Gene enrichment analysis for disPCA without the HLA region.
Table shows pathways that are enriched in the *disPCA* analysis based on the GSEA analysis after removing genes in the HLA and surrounding region.

Pairs of datasets of the same disease			
PC	**Ranked by**	**Correlation**	**p-value**
1	Physical	0.62	1.7×10^{-6}
	Genetic	0.69	3.6×10^{-8}
2	Physical	0.51	1.0×10^{-4}
	Genetic	0.31	0.0287
mean(PC1,PC2)	Physical	0.74	1.1×10^{-9}
	Genetic	0.67	8.4×10^{-8}

Table 4.S1. Comparison of loadings between disPCA with mapping based on physical or genetic coordinates.
Loadings for the top 50 genes ranked by either a physical or genetic coordinates based *disPCA* were compared. 'Correlation' denotes the Pearson's correlation coefficient with its significance denoted in the 'p-value' column. Rows denoted by 'mean(PC1,PC2)' indicate the correlation between the 50 genes with the largest average loading of PC1 and PC2.

Study Name	Disease	Obtained via	Association Method	Array	Sample Size
ALS Finland(Laaksovirta, Peuralinna et al. 2010) *(ALS_Fin)*	ALS	dbGaP	Logistic regression	Overlap between Illumina 1M and CNV 370	973
ALS Irish(Cronin, Berger et al. 2008)*(ALS_Irish)*	ALS	dbGaP	Logistic regression	Illumina 550k	432
Duke Alzheimer's(Heinzen, Need et al. 2010) *(Alz_Duke)*	Alzheimer's Disease	http://humangenome.duke.edu/available-datasets	Logistic regression	Illumina 550	699
GenADA *(Li, Wetten et al. 2008)(Alz_GenADA)*	Alzheimer's Disease	dbGaP	Logistic regression	Affymetrix 400	1588
WTCCC2 AS(Evans, Spencer et al. 2011) (*AS_WT)*	Ankyolosing Spondylitis	WTCCC	Logistic regression	Illumina 1M	2732
ADHD PGC(Neale, Medland et al. 2010) (*ADHD_PGC)*	ADHD	PGC	Meta-analysis	Imputation	5415
Behcet's GWAS(Remmers, Cosan et al. 2010) *(Behcets_GWAS)*	Bechet's	dbGaP	Chis-sq	Illumina CNV 370	2493
BMI Stampeed (Sabatti, Service et al. 2009)(*BMI_Stampeed)*	BMI	dbGaP	Linear regression	Illumina CNV 370	5415
BMI Sardinia(Scuteri, Sanna et al. 2007) (*BMI_Sardin)*	BMI	dbGaP	Merlin	Affymetrix 500	1412
CGEMS Breast Cancer(Hunter, Kraft et al. 2007) *(BreastC_CGEMS)*	Breast Cancer	dbGaP	Logistic regression	Illumina 550	2287
CIDR Celiac(Ahn, Ding et al. 2012) *(CeliacD_CIDR)*	Celiac disease	dbGaP	Logistic regression	Illumina 660	2246
NIDDK IBD(Duerr, Taylor et al. 2006) *(CD_NIDDK)*	Crohn's disease	dbGaP	Chis-sq	Illumina 300	1028
WTCCC CD(The Wellcome Trust Case Control Consortium 2007) *(CD_WTCCC)*	Crohn's disease	WTCCC	Logistic regression	Affymetrix 500	3293
Ischemic Stroke(Matarin, Brown et al. 2007) (*IscStroke)*	Ischemic Stroke	dbGaP	Logistic regression	Illumina 300	485
Major Depression GWAS(Boomsma,	Major depression	dbGaP	Logistic regression	Perlgen 600k	3741

Willemsen et al. 2008) (*MajDep)*					
WTCCC2 MS(Sawcer, Hellenthal et al. 2011) (*MS_WT)*	Multiple Sclerosis	WTCCC	Logistic regression	Illumina 1M	4055
GeneMSA(Baranzini, Wang et al. 2009) *(MS_GeneMSA)*	Multiple Sclerosis	dbGaP	Logistic Regression	Illumina 550	2000
CIDR Parkinson's(Karamoham ed, Golbe et al. 2005; Nichols, Pankratz et al. 2005) *(Parkin_CIDR)*	Parkinson's	dbGaP	Logistic regression	Illumina CNV 370	1991
CASP(Helms, Cao et al. 2003; Nair, Stuart et al. 2006; Nair, Duffin et al. 2009) *(Psor_CASP)*	Psoriasis	dbGaP	Chi-sq	Perlgen 600k	2825
WTCCC RA (The Wellcome Trust Case Control Consortium 2007) *(RA_WTCCC)*	Rheumatoid arthritis	WTCCC	Logistic regression	Affymetrix 500	3481
Schizophrenia GWAS(Suarez, Duan et al. 2006) *(Schizo_GWAS)*	Schizophrenia	dbGaP	Chi-sq	Affymetrix 6.0	2659
PGC Schizophrenia(Schizoph renia Psychiatric Genome-Wide Association Study (GWAS) Consortium 2011) *(Schizo_PGC)*	Schizophrenia	PGC	Meta-analysis	Imputation	21,856
SLEGEN(Harley, Alarcon-Riquelme et al. 2008) *(SLE_SLEGEN)*	SLE	dbGaP	additive model	Illumina 300	297
SLE GWAS(Hom, Graham et al. 2008) *(SLE_GWAS)*	SLE	dbGaP	Chi-sq	Illumina 550	4651
T2D Fusion(Scott, Mohlke et al. 2007) *(T2D_Fusion)*	T2D	dbGaP	Logistic regression	Illumina 300	1706
T2D Scandinavia(Saxena, Voight et al. 2007) *(T2D_Scandinavia)*	T2D	http://www.broad institute.org/diabe tes/scandinavs/typ e2.html	Cochran-Mantel-Haenszel	Affymetrix 500	3000
WTCCC2 UC(Barrett, Lee et al. 2009) (*UC_WT)*	Ulcerative colitis	WTCCC	Logistic regression	Affymetrix 6.0	404
VitGene(Jin, Birlea et al. 2010) *(Vit_GWAS1)*	Vitiligo	dbGaP	Logistic regression	Illumina 610	4327
Vitiligo GWAS2(Jin, Birlea et al. 2012) *(Vit_GWAS2)*	Vitiligo	Collaboration	Logistic regression	Illumina 660	3632

Table 4.S2. Dataset attributes.
Various attributes of datasets utilized in this study.

Replication			
PC	**Ranked by**	**Correlation**	**p-value**
1	Replication 1	-0.056	0.7
	Replication 2	0.28	0.049
2	Replication 1	0.479	4.3×10^{-4}
	Replication 2	0.634	7.7×10^{-7}
mean(PC1,PC2)	Replication 1	0.444	1.2×10^{-3}
	Replication 2	0.652	7.7×10^{-7}

Table 4.S3. Comparison of loadings between Replication Sets 1 and 2.
Loadings for the top 50 genes ranked by either Replication Set 1 or Replication Set 2 were compared. 'Correlation' denotes the Pearson's correlation coefficient with its significance denoted in the 'p-value' column. Rows denoted by 'mean(PC1,PC2)' indicate the correlation between the 50 genes with the largest average loading of PC1 and PC2.

PC	Pathway	FDR (q-value)
1	Intestinal immune network for IgA production	0.028
	Antigen processing and presentation	0.057
	Spliceosome	0.141
	Inositol phosphate metabolism	0.156
	Cell adhesion molecules	0.19
2	NOD-like receptor signaling pathway	0.152
	GH pathway	0.207
	Insulin pathway	0.211
	CardiacEGF pathway	0.213
	IL2 pathway	0.226
	NFAT pathway	0.236
	Dorso ventral axis formation	0.236
	IL2RB pathway	0.245
	IGF-1 pathway	0.246

Table 4.S4. Pathway enrichment after filtering nearby genes.
Pathway enrichment was applied to a subset of genes that were located greater than 0.1cM from each other

References

Abdou, A. G., A. H. Maraee, et al. (2013). "Immunohistochemical expression of heat shock protein 70 in vitiligo." Annals of diagnostic pathology **17**(3): 245-249.

Abecasis, G. R., A. Auton, et al. (2012). "An integrated map of genetic variation from 1,092 human genomes." Nature **491**(7422): 56-65.

Adams, A. M. and R. R. Hudson (2004). "Maximum-likelihood estimation of demographic parameters using the frequency spectrum of unlinked single-nucleotide polymorphisms." Genetics **168**: 1699-1712.

Ahmeti, K. B., S. Ajroud-Driss, et al. (2013). "Age of onset of amyotrophic lateral sclerosis is modulated by a locus on 1p34.1." Neurobiology of Aging **34**(1): 357 e357-319.

Ahn, R., Y. C. Ding, et al. (2012). "Association analysis of the extended MHC region in celiac disease implicates multiple independent susceptibility loci." PloS one **7**(5): e36926.

Albrechtsen, A., F. C. Nielsen, et al. (2010). "Ascertainment biases in SNP chips affect measures of population divergence." Molecular biology and evolution **27**(11): 2534-2547.

Aleman, A., R. S. Kahn, et al. (2003). "Sex differences in the risk of schizophrenia: evidence from meta-analysis." Archives of general psychiatry **60**(6): 565-571.

Altshuler, D. M., R. A. Gibbs, et al. (2010). "Integrating common and rare genetic variation in diverse human populations." Nature **467**: 52-58.

Amberger, J., C. Bocchini, et al. (2011). "A new face and new challenges for Online Mendelian Inheritance in Man (OMIM(R))." Human Mutation **32**(5): 564-567.

Amberger, J., C. A. Bocchini, et al. (2009). "McKusick's Online Mendelian Inheritance in Man (OMIM)." Nucleic Acids Research **37**(Database issue): D793-796.

Amundadottir, L., P. Kraft, et al. (2009). "Genome-wide association study identifies variants in the ABO locus associated with susceptibility to pancreatic cancer." Nature Genetics **41**(9): 986-990.

Andersen, K., L. J. Launer, et al. (1999). "Gender differences in the incidence of AD and vascular dementia: The EURODEM Studies. EURODEM Incidence Research Group." Neurology **53**(9): 1992-1997.

Anderson, C. A., N. Soranzo, et al. (2011). "Synthetic associations are unlikely to account for many common disease genome-wide association signals." PLoS Biology **9**(1): e1000580.

Anderson, K. M., P. M. Odell, et al. (1991). "Cardiovascular disease risk profiles." American heart journal **121**(1 Pt 2): 293-298.

Andreassen, O. A., H. F. Harbo, et al. (2014). "Genetic pleiotropy between multiple sclerosis and schizophrenia but not bipolar disorder: differential involvement of immune-related gene loci." Molecular Psychiatry.

Andreassen, O. A., W. K. Thompson, et al. (2013). "Improved detection of common variants associated with schizophrenia and bipolar disorder using pleiotropy-informed conditional false discovery rate." PLoS Genetics **9**(4): e1003455.

Ashburner, M., C. A. Ball, et al. (2000). "Gene ontology: tool for the unification of biology. The Gene Ontology Consortium." Nature Genetics **25**(1): 25-29.

Baek, H. Y., J. W. Lim, et al. (2007). "Interaction between the Helicobacter pylori CagA and alpha-Pix in gastric epithelial AGS cells." Annals of the New York Academy of Sciences **1096**: 18-23.

Bansal, V., O. Libiger, et al. (2010). "Statistical analysis strategies for association studies involving rare variants." Nature Reviews Genetics **11**(11): 773-785.

Baranzini, S. E., J. Wang, et al. (2009). "Genome-wide association analysis of susceptibility and clinical phenotype in multiple sclerosis." Human Molecular Genetics **18**(4): 767-778.

Baron-Cohen, S., M. V. Lombardo, et al. (2011). "Why are autism spectrum conditions more prevalent in males?" PLoS Biology **9**(6): e1001081.

Barrett, J. C., D. G. Clayton, et al. (2009). "Genome-wide association study and meta-analysis find that over 40 loci affect risk of type 1 diabetes." Nature Genetics **41**(6): 703-707.

Barrett, J. C., J. C. Lee, et al. (2009). "Genome-wide association study of ulcerative colitis identifies three new susceptibility loci, including the HNF4A region." Nature Genetics **41**(12): 1330-1334.

Barrett, T., S. E. Wilhite, et al. (2013). "NCBI GEO: archive for functional genomics data sets--update." Nucleic Acids Research **41**(Database issue): D991-995.

Beeson, P. B. (1994). "Age and sex associations of 40 autoimmune diseases." The American journal of medicine **96**(5): 457-462.

Bellamy, R., N. Beyers, et al. (2000). "Genetic susceptibility to tuberculosis in Africans: a genome-wide scan." Proceedings of the National Academy of Sciences of the United States of America **97**: 8005-8009.

Bennett, C. L., J. Christie, et al. (2001). "The immune dysregulation, polyendocrinopathy, enteropathy, X-linked syndrome (IPEX) is caused by mutations of FOXP3." Nature Genetics **27**(1): 20-21.

Benros, M. E., W. W. Eaton, et al. (2013). "The epidemiological evidence linking autoimmune diseases and psychosis." Biological Psychiatry.

Berndt, S. I., S. Gustafsson, et al. (2013). "Genome-wide meta-analysis identifies 11 new loci for anthropometric traits and provides insights into genetic architecture." Nature Genetics

45(5): 501-512.

Beyene, J., D. Tritchler, et al. (2009). "Gene- or region-based analysis of genome-wide association studies." Genetic Epidemiology **33 Suppl 1**: S105-110.

Bianchi, I., A. Lleo, et al. (2011). "The X chromosome and immune associated genes." Journal of autoimmunity.

Bierut, L. J. (2007). "Genetic variation that contributes to nicotine dependence." Pharmacogenomics **8**(8): 881-883.

Bierut, L. J., N. L. Saccone, et al. (2002). "Defining alcohol-related phenotypes in humans. The Collaborative Study on the Genetics of Alcoholism." Alcohol research & health : the journal of the National Institute on Alcohol Abuse and Alcoholism **26**(3): 208-213.

Bierut, L. J., J. R. Strickland, et al. (2008). "Drug use and dependence in cocaine dependent subjects, community-based individuals, and their siblings." Drug and Alcohol Dependence **95**(1-2): 14-22.

Birlea, S. A., Y. Jin, et al. (2011). "Comprehensive association analysis of candidate genes for generalized vitiligo supports XBP1, FOXP3, and TSLP." The Journal of investigative dermatology **131**(2): 371-381.

Bodmer, W. and C. Bonilla (2008). "Common and rare variants in multifactorial susceptibility to common diseases." Nature Genetics **40**(6): 695-701.

Bonen, D. K. and J. H. Cho (2003). "The genetics of inflammatory bowel disease." Gastroenterology **124**: 521-536.

Boomsma, D. I., G. Willemsen, et al. (2008). "Genome-wide association of major depression: description of samples for the GAIN Major Depressive Disorder Study: NTR and NESDA biobank projects." European journal of human genetics : EJHG **16**(3): 335-342.

Bouvet, J. P. and V. A. Fischetti (1999). "Diversity of antibody-mediated immunity at the mucosal barrier." Infection and Immunity **67**(6): 2687-2691.

Broadley, S. A., J. Deans, et al. (2000). "Autoimmune disease in first-degree relatives of patients with multiple sclerosis. A UK survey." Brain : a journal of neurology **123 (Pt 6)**: 1102-1111.

Browning, S. R. and B. L. Browning (2007). "Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering." American Journal of Human Genetics **81**(5): 1084-1097.

Buckley, P. F., S. Mahadik, et al. (2007). "Neurotrophins and schizophrenia." Schizophrenia Research **94**(1-3): 1-11.

Calvo, J. R., C. Gonzalez-Yanes, et al. (2013). "The role of melatonin in the cells of the innate immunity: a review." Journal of Pineal Research **55**(2): 103-120.

Carrel, L. and H. F. Willard (2005). "X-inactivation profile reveals extensive variability in X-linked gene expression in females." Nature **434**(7031): 400-404.

Chang, D., F. Gao, et al. "XWAS: a toolset for genetic data analysis and association studies of the X chromosome." Under Review.

Chang, D., F. Gao, et al. (2014). "Accounting for eXentricities: Analysis of the X chromosome in GWAS reveals X-linked genes implicated in autoimmune diseases." Submitted.

Chang, D. and A. Keinan (2012). "Predicting signatures of "synthetic associations" and "natural associations" from empirical pattherns of human genetic variation " PLos Computational Biology **8**(7): e1002600.

Chang, D. and A. Keinan (2014). "Principal component analysis characterizes shared pathogenetics from genome-wide association studies." in revision.

Chapman, J. M., J. D. Cooper, et al. (2003). "Detecting disease associations due to linkage disequilibrium using haplotype tags: a class of tests and the determinants of statistical power." Human Heredity **56**(1-3): 18-31.

Chataway, J., R. Feakes, et al. (1998). "The genetics of multiple sclerosis: principles, background and updated results of the United Kingdom systematic genome screen." Brain : a journal of neurology **121 (Pt 10)**: 1869-1887.

Choi, B. G. and M. A. McLaughlin (2007). "Why men's hearts break: cardiovascular effects of sex steroids." Endocrinology and metabolism clinics of North America **36**(2): 365-377.

Cirulli, E. T. and D. B. Goldstein (2010). "Uncovering the roles of rare variants in common disease through whole-genome sequencing." Nature Reviews Genetics **11**(6): 415-425.

Clark, A. G., M. J. Hubisz, et al. (2005). "Ascertainment bias in studies of human genome-wide polymorphism." Genome Research **15**(11): 1496-1502.

Clayton, D. (2008). "Testing for association on the X chromosome." Biostatistics **9**(4): 593-600.

Clayton, D. G. (2009). "Sex chromosomes and genetic association studies." Genome medicine **1**: 110.

Cohen, J. C., E. Boerwinkle, et al. (2006). "Sequence variations in PCSK9, low LDL, and protection against coronary heart disease." The New England Journal of Medicine **354**: 1264-1272.

Conde, L., J. N. Foo, et al. (2013). "X chromosome-wide association study of follicular lymphoma." British Journal of Haematology **162**(6): 858-862.

Confavreux, C., M. Hutchinson, et al. (1998). "Rate of pregnancy-related relapse in multiple sclerosis. Pregnancy in Multiple Sclerosis Group." The New England Journal of Medicine **339**(5): 285-291.

Cotsapas, C., B. F. Voight, et al. (2011). "Pervasive sharing of genetic effects in autoimmune disease." PLoS Genetics **7**(8): e1002254.

Coventry, A., L. M. Bull-Otterson, et al. (2010). "Deep resequencing reveals excess rare recent variants consistent with explosive population growth." Nature Communications **1**: 131.

Cronin, S., S. Berger, et al. (2008). "A genome-wide association study of sporadic ALS in a homogenous Irish population." Human Molecular Genetics **17**(5): 768-774.

Cunningham-Rundles, C. (2001). "Physiology of IgA and IgA deficiency." Journal of Clinical Immunology **21**(5): 303-309.

Darabos, C., K. Desai, et al. (2013). Evolutionary Computation, Machine Learning and Data Mining in Bioinformatics, Springer Berlin Heidelberg.

Dibner, C., U. Schibler, et al. (2010). "The Mammalian Circadian Timing System: Organization and Coordination of Central and Peripheral Clocks." Annual Review of Physiology **72**: 517-549.

Dickson, S. P., K. Wang, et al. (2010). "Rare variants create synthetic genome-wide associations." PLoS Biology **8**: e1000294.

Dobyns, W. B., A. Filauro, et al. (2004). "Inheritance of most X-linked traits is not dominant or recessive, just X-linked." American journal of medical genetics. Part A **129A**(2): 136-143.

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

Dunning, A. M., F. Durocher, et al. (2000). "The extent of linkage disequilibrium in four populations with distinct demographic histories." American Journal of Human Genetics **67**(6): 1544-1554.

Durany, N., T. Michel, et al. (2001). "Brain-derived neurotrophic factor and neurotrophin 3 in schizophrenic psychoses." Schizophrenia Research **52**(1-2): 79-86.

Eguchi, K. (2001). "Apoptosis in autoimmune diseases." Internal medicine **40**(4): 275-284.

Eichler, E. E., J. Flint, et al. (2010). "Missing heritability and strategies for finding the underlying causes of complex disease." Nature Reviews Genetics **11**(6): 446-450.

Ellegren, H. and J. Parsch (2007). "The evolution of sex-biased genes and sex-biased gene expression." Nature reviews. Genetics **8**(9): 689-698.

Ellinghaus, D., E. Ellinghaus, et al. (2012). "Combined analysis of genome-wide association studies for Crohn disease and psoriasis identifies seven shared susceptibility loci." American Journal of Human Genetics **90**(4): 636-647.

Estrada, K., U. Styrkarsdottir, et al. (2012). "Genome-wide meta-analysis identifies 56 bone mineral density loci and reveals 14 loci associated with risk of fracture." Nature Genetics **44**(5): 491-501.

Evans, D. M., C. C. Spencer, et al. (2011). "Interaction between ERAP1 and HLA-B27 in ankylosing spondylitis implicates peptide handling in the mechanism for HLA-B27 in disease susceptibility." Nature Genetics **43**(8): 761-767.

Fellay, J., A. J. Thompson, et al. (2010). "ITPA gene variants protect against anaemia in patients treated for chronic hepatitis C." Nature **464**(7287): 405-408.

Festen, E. A., P. Goyette, et al. (2011). "A meta-analysis of genome-wide association scans identifies IL18RAP, PTPN2, TAGAP, and PUS10 as shared risk loci for Crohn's disease and celiac disease." PLoS Genet **7**(1): e1001283.

Fish, E. N. (2008). "The X-files in immunity: sex-based differences predispose immune responses." Nature reviews. Immunology **8**(9): 737-744.

Fisher, R. A. (1925). Statistical Methods for Research Workers. Edinburgh, Oliver and Boyd.

Fontenot, J. D., M. A. Gavin, et al. (2003). "Foxp3 programs the development and function of CD4(+)CD25(+) regulatory T cells." Nature Immunology **4**(4): 330-336.

Frazer, K. A., D. G. Ballinger, et al. (2007). "A second generation human haplotype map of over 3.1 million SNPs." Nature **449**: 851-861.

Frazer, K. A., S. S. Murray, et al. (2009). "Human genetic variation and its contribution to complex traits." Nature Reviews Genetics **10**(4): 241-251.

Gater, R., M. Tansella, et al. (1998). "Sex differences in the prevalence and detection of depressive and anxiety disorders in general health care settings: report from the World Health Organization Collaborative Study on Psychological Problems in General Health Care." Archives of general psychiatry **55**(5): 405-413.

Gleicher, N. and D. H. Barad (2007). "Gender as risk factor for autoimmune diseases." Journal of Autoimmunity **28**(1): 1-6.

Goh, K. I., M. E. Cusick, et al. (2007). "The human disease network." Proceedings of the National Academy of Sciences of the United States of America **104**(21): 8685-8690.

Goldstein, D. B. (2011). "The Importance of Synthetic Associations Will Only Be Resolved Empirically." PLoS Biology **9**: e1001008.

Goldstein, J. M., L. J. Seidman, et al. (2001). "Normal sexual dimorphism of the adult human brain assessed by in vivo magnetic resonance imaging." Cerebral Cortex **11**(6): 490-497.

Gottipati, S., L. Arbiza, et al. (2011). "Analyses of X-linked and autosomal genetic variation in population-scale whole genome sequencing." Nature Genetics **43**(8): 741-743.

Gray, K. A., L. C. Daugherty, et al. (2013). "Genenames.org: the HGNC resources in 2013." Nucleic Acids Research **41**(Database issue): D545-552.

Green, E. D. and M. S. Guyer (2011). "Charting a course for genomic medicine from base pairs to bedside." Nature **470**(7333): 204-213.

Gusev, A., G. Bhatia, et al. (2013). "Quantifying missing heritability at known GWAS loci." PLoS Genetics **9**(12): e1003993.

Hakonarson, H., S. F. Grant, et al. (2007). "A genome-wide association study identifies KIAA0350 as a type 1 diabetes gene." Nature **448**(7153): 591-594.

Hamdouch, K., C. Rodriguez, et al. (2011). "Anti-CENPI autoantibodies in scleroderma patients with features of autoimmune liver diseases." Clinica chimica acta; international journal of clinical chemistry **412**(23-24): 2267-2271.

Hammer, M. F., F. L. Mendez, et al. (2008). "Sex-biased evolutionary forces shape genomic patterns of human diversity." PLoS Genetics **4**(9): e1000202.

Hammer, M. F., A. E. Woerner, et al. (2010). "The ratio of human X chromosome to autosome diversity is positively correlated with genetic distance from genes." Nature Genetics **42**(10): 830-831.

Hamosh, A., A. F. Scott, et al. (2002). "Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders." Nucleic Acids Research **30**(1): 52-55.

Hamosh, A., A. F. Scott, et al. (2005). "Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders." Nucleic Acids Research **33**(Database issue): D514-517.

Han, F. and W. Pan (2010). "A data-adaptive sum test for disease association with multiple common or rare variants." Human heredity **70**(1): 42-54.

Harley, J. B., M. E. Alarcon-Riquelme, et al. (2008). "Genome-wide association scan in women

with systemic lupus erythematosus identifies susceptibility variants in ITGAM, PXK, KIAA1542 and other loci." Nature Genetics **40**(2): 204-210.

Harper, P. S. (1984). Practical genetic counselling, Wright.

Hartley, S. W., S. Monti, et al. (2012). "Bayesian methods for multivariate modeling of pleiotropic SNP associations and genetic risk prediction." Frontiers in genetics **3**: 176.

Heinzen, E. L., A. C. Need, et al. (2010). "Genome-wide scan of copy number variation in late-onset Alzheimer's disease." Journal of Alzheimer's disease : JAD **19**(1): 69-77.

Heller, M. M., E. S. Lee, et al. (2011). "Stress as an influencing factor in psoriasis." Skin therapy letter **16**(5): 1-4.

Helms, C., L. Cao, et al. (2003). "A putative RUNX1 binding site variant between SLC9A3R1 and NAT9 is associated with susceptibility to psoriasis." Nature Genetics **35**(4): 349-356.

Heyer, E., R. Chaix, et al. (2012). "Sex-specific demographic behaviours that shape human genomic variation." Molecular Ecology **21**(3): 597-612.

Hindorff, L. A., J. MacArthur, et al. (2013). A Catalog of Published Genome-wide Association Studies.

Hindorff, L. A., P. Sethupathy, et al. (2009). "Potential etiologic and functional implications of genome-wide association loci for human diseases and traits." Proceedings of the National Academy of Sciences of the United States of America **106**(23): 9362-9367.

Hoffmann, T. J., N. J. Marini, et al. (2010). "Comprehensive Approach to Analyzing Rare Genetic Variants." PLoS ONE **5**: e13584.

Hofman, A., M. M. Breteler, et al. (2009). "The Rotterdam Study: 2010 objectives and design update." European Journal of Epidemiology **24**(9): 553-572.

Hom, G., R. R. Graham, et al. (2008). "Association of systemic lupus erythematosus with

C8orf13-BLK and ITGAM-ITGAX." The New England Journal of Medicine **358**(9): 900-909.

Howie, B. N., P. Donnelly, et al. (2009). "A flexible and accurate genotype imputation method for the next generation of genome-wide association studies." PLoS Genetics **5**(6): e1000529.

Huang, H. L., P. Chanda, et al. (2011). "Gene-Based Tests of Association." PLoS Genetics **7**(7).

Hugot, J. P., M. Chamaillard, et al. (2001). "Association of NOD2 leucine-rich repeat variants with susceptibility to Crohn's disease." Nature **411**(6837): 599-603.

Hunter, D. J., P. Kraft, et al. (2007). "A genome-wide association study identifies alleles in FGFR2 associated with risk of sporadic postmenopausal breast cancer." Nature Genetics **39**(7): 870-874.

Huynh, K. D., W. Fischle, et al. (2000). "BCoR, a novel corepressor involved in BCL-6 repression." Genes & Development **14**(14): 1810-1823.

Iles, M. M. (2008). "What can genome-wide association studies tell us about the genetics of common disease?" PLoS genetics **4**(2): e33.

Itariu, B. K. and T. M. Stulnig (2014). "Autoimmune Aspects of Type 2 Diabetes Mellitus - A Mini-Review." Gerontology.

Jacob, S., B. Poeggeler, et al. (2002). "Melatonin as a candidate compound for neuroprotection in amyotrophic lateral sclerosis (ALS): high tolerability of daily oral melatonin administration in ALS patients." Journal of Pineal Research **33**(3): 186-187.

Jazin, E. and L. Cahill (2010). "Sex differences in molecular neuroscience: from fruit flies to humans." Nature reviews. Neuroscience **11**(1): 9-17.

Jiang, B., X. Zhang, et al. (2011). "A powerful truncated tail strength method for testing multiple

null hypotheses in one dataset." Journal of Theoretical Biology **277**(1): 67-73.

Jin, X., Y. P. Chen, et al. (2013). "Association between Helicobacter Pylori infection and ulcerative colitis--a case control study from China." International journal of medical sciences **10**(11): 1479-1484.

Jin, Y., S. A. Birlea, et al. (2012). "Genome-wide association analyses identify 13 new susceptibility loci for generalized vitiligo." Nature Genetics **44**(6): 676-680.

Jin, Y., S. A. Birlea, et al. (2010). "Variant of TYR and autoimmunity susceptibility loci in generalized vitiligo." The New England Journal of Medicine **362**(18): 1686-1697.

Jorgenson, E. and J. S. Witte (2006). "A gene-centric approach to genome-wide association studies." Nature reviews. Genetics **7**(11): 885-891.

Jostins, L., S. Ripke, et al. (2012). "Host-microbe interactions have shaped the genetic architecture of inflammatory bowel disease." Nature **491**(7422): 119-124.

Ju, T. and R. D. Cummings (2005). "Protein glycosylation: chaperone mutation in Tn syndrome." Nature **437**(7063): 1252.

Kanehisa, M. and S. Goto (2000). "KEGG: kyoto encyclopedia of genes and genomes." Nucleic Acids Research **28**(1): 27-30.

Karamohamed, S., L. I. Golbe, et al. (2005). "Absence of previously reported variants in the SCNA (G88C and G209A), NR4A2 (T291D and T245G) and the DJ-1 (T497C) genes in familial Parkinson's disease from the GenePD study." Movement disorders : official journal of the Movement Disorder Society **20**(9): 1188-1191.

Kathiresan, S., C. J. Willer, et al. (2009). "Common variants at 30 loci contribute to polygenic dyslipidemia." Nature Genetics **41**(1): 56-65.

Kawakami, A. and K. Eguchi (2002). "Involvement of apoptotic cell death in autoimmune

diseases." Medical electron microscopy : official journal of the Clinical Electron Microscopy Society of Japan **35**(1): 1-8.

Keinan, A. and A. G. Clark (2012). "Recent explosive population growth has resulted in massive excess of rare variants in humans." Science **In press**.

Keinan, A., J. C. Mullikin, et al. (2007). "Measurement of the human allele frequency spectrum demonstrates greater genetic drift in East Asians than in Europeans." Nature Genetics **39**: 1251-1255.

Keinan, A., J. C. Mullikin, et al. (2009). "Accelerated genetic drift on chromosome X during the human dispersal out of Africa." Nature Genetics **41**: 66-70.

Keinan, A. and D. Reich (2010). "Can a sex-biased human demography account for the reduced effective population size of chromosome X in non-Africans?" Molecular Biology and Evolution **27**(10): 2312-2321.

Kemkemer, C., M. Kohn, et al. (2009). "Enrichment of brain-related genes on the mammalian X chromosome is ancient and predates the divergence of synapsid and sauropsid lineages." Chromosome research : an international journal on the molecular, supramolecular and evolutionary aspects of chromosome biology **17**(6): 811-820.

Kendler, K. S. and S. R. Diehl (1993). "The genetics of schizophrenia: a current, genetic-epidemiologic perspective." Schizophrenia Bulletin **19**(2): 261-285.

Klei, L., D. Luca, et al. (2008). "Pleiotropy and principal components of heritability combine to increase power for association analysis." Genetic Epidemiology **32**(1): 9-19.

Konig, I. R., C. Loley, et al. (2014). "How to include chromosome x in your genome-wide association study." Genetic Epidemiology **38**(2): 97-103.

Korte, A., B. J. Vilhjalmsson, et al. (2012). "A mixed-model approach for genome-wide

association studies of correlated traits in structured populations." Nature Genetics **44**(9): 1066-1071.

Laaksovirta, H., T. Peuralinna, et al. (2010). "Chromosome 9p21 in amyotrophic lateral sclerosis in Finland: a genome-wide association study." Lancet neurology **9**(10): 978-985.

Lai, F., E. Kammann, et al. (1999). "APOE genotype and gender effects on Alzheimer disease in 100 adults with Down syndrome." Neurology **53**(2): 331-336.

Lee, J. H., R. Cheng, et al. (2008). "Analyses of the National Institute on Aging Late-Onset Alzheimer's Disease Family Study: implication of additional loci." Archives of Neurology **65**(11): 1518-1526.

Lee, P. H., S. E. Bergen, et al. (2011). "Modifiers and subtype-specific analyses in whole-genome association studies: a likelihood framework." Human Heredity **72**(1): 10-20.

Lee, S. H., S. Ripke, et al. (2013). "Genetic relationship between five psychiatric disorders estimated from genome-wide SNPs." Nature Genetics **45**(9): 984-994.

Lerner, D. J. and W. B. Kannel (1986). "Patterns of coronary heart disease morbidity and mortality in the sexes: a 26-year follow-up of the Framingham population." American heart journal **111**(2): 383-390.

Li, H., S. Wetten, et al. (2008). "Candidate single-nucleotide polymorphisms from a genomewide association study of Alzheimer disease." Archives of Neurology **65**(1): 45-53.

Li, M. X., H. S. Gui, et al. (2011). "GATES: a rapid and powerful gene-based association test using extended Simes procedure." American Journal of Human Genetics **88**(3): 283-293.

Li, N., T. Ma, et al. (2014). "Increased apoptosis induction in CD4+CD25+ Foxp3+ T cells contributes to enhanced disease activity in patients with rheumatoid arthritis through Il-10 regulation." European review for medical and pharmacological sciences **18**(1): 78-85.

Libert, C., L. Dejager, et al. (2010). "The X chromosome in immune functions: when a chromosome makes the difference." Nature reviews. Immunology **10**(8): 594-604.

Liu, J. Z., A. F. McRae, et al. (2010). "A versatile gene-based test for genome-wide association studies." American Journal of Human Genetics **87**(1): 139-145.

Lockshin, M. D. (2006). "Sex differences in autoimmune disease." Lupus **15**(11): 753-756.

Logan, C. Y. and R. Nusse (2004). "The Wnt signaling pathway in development and disease." Annual Review of Cell and Developmental Biology **20**: 781-810.

Lohmueller, K. E., J. D. Degenhardt, et al. (2010). "Sex-averaged recombination and mutation rates on the X chromosome: a comment on Labuda et al." American Journal of Human Genetics **86**(6): 978-980; author reply 980-971.

Longmate, J. A., G. P. Larson, et al. (2010). "Three ways of combining genotyping and resequencing in case-control association studies." PloS ONE **5**: e14318.

Ludwig, D., M. Stahl, et al. (1999). "Enhanced intestinal expression of heat shock protein 70 in patients with inflammatory bowel diseases." Digestive Diseases and Sciences **44**(7): 1440-1447.

Lukashev, M. E. and Z. Werb (1998). "ECM signalling: orchestrating cell behaviour and misbehaviour." Trends in Cell Biology **8**(11): 437-441.

Luna, A. and K. K. Nicodemus (2007). "snp.plotter: an R-based SNP/haplotype association and linkage disequilibrium plotting package." Bioinformatics **23**(6): 774-776.

Luther, J., M. Dave, et al. (2010). "Association between Helicobacter pylori infection and inflammatory bowel disease: a meta-analysis and systematic review of the literature." Inflammatory bowel diseases **16**(6): 1077-1084.

Ma, L., A. G. Clark, et al. (2013). "Gene-Based Testing of Interactions in Association Studies of

Quantitative Traits." PLoS Genetics **9**(2).

Macpherson, A., U. Y. Khoo, et al. (1996). "Mucosal antibodies in inflammatory bowel disease are directed against intestinal bacteria." Gut **38**(3): 365-375.

Madsen, B. E. and S. R. Browning (2009). "A groupwise association test for rare mutations using a weighted sum statistic." PLoS genetics **5**(2): e1000384.

Mahan, A. L. and K. J. Ressler (2012). "Fear conditioning, synaptic plasticity and the amygdala: implications for posttraumatic stress disorder." Trends in Neurosciences **35**(1): 24-35.

Maher, B. (2008). "Personal genomes: The case of the missing heritability." Nature **456**(7218): 18-21.

Manolio, T. A., F. S. Collins, et al. (2009). "Finding the missing heritability of complex diseases." Nature **461**(7265): 747-753.

Marigorta, U. M. and A. Navarro (2013). "High Trans-ethnic Replicability of GWAS Results Implies Common Causal Variants." PLoS Genetics **9**(6): e1003566.

Marrie, R. A., R. I. Horwitz, et al. (2011). "Smokers with multiple sclerosis are more likely to report comorbid autoimmune diseases." Neuroepidemiology **36**(2): 85-90.

Marth, G. T., E. Czabarka, et al. (2004). "The allele frequency spectrum in genome-wide human variation data reveals signals of differential demographic history in three large world populations." Genetics **166**: 351-372.

Mason, K. D., A. Lin, et al. (2013). "Proapoptotic Bak and Bax guard against fatal systemic and organ-specific autoimmune disease." Proceedings of the National Academy of Sciences of the United States of America **110**(7): 2599-2604.

Matanoski, G., X. Tao, et al. (2006). "Demographics and tumor characteristics of colorectal cancers in the United States, 1998-2001." Cancer **107**(5 Suppl): 1112-1120.

Matarin, M., W. M. Brown, et al. (2007). "A genome-wide genotyping study in patients with ischaemic stroke: initial analysis and data release." Lancet neurology **6**(5): 414-420.

Matson, D. R., P. B. Demirel, et al. (2012). "A conserved role for COMA/CENP-H/I/N kinetochore proteins in the spindle checkpoint." Genes & Development **26**(6): 542-547.

McCarthy, M. I., G. R. Abecasis, et al. (2008). "Genome-wide association studies for complex traits: consensus, uncertainty and challenges." Nature Reviews Genetics **9**(5): 356-369.

McCarthy, M. I. and J. N. Hirschhorn (2008). "Genome-wide association studies: potential next steps on a genetic journey." Human Molecular Genetics **17**: R156-165.

McElroy, S. L. (2004). "Diagnosing and treating comorbid (complicated) bipolar disorder." The Journal of clinical psychiatry **65 Suppl 15**: 35-44.

Mendelsohn, M. E. and R. H. Karas (2005). "Molecular and cellular basis of cardiovascular gender differences." Science **308**(5728): 1583-1587.

Momose, Y., M. Murata, et al. (2002). "Association studies of multiple candidate genes for Parkinson's disease using single nucleotide polymorphisms." Annals of Neurology **51**(1): 133-136.

Mosenson, J. A., J. M. Eby, et al. (2013). "A central role for inducible heat-shock protein 70 in autoimmune vitiligo." Experimental dermatology **22**(9): 566-569.

Mosenson, J. A., A. Zloza, et al. (2012). "HSP70i is a critical component of the immune response leading to vitiligo." Pigment cell & melanoma research **25**(1): 88-98.

Muscat, J. E., J. P. Richie, Jr., et al. (1996). "Gender differences in smoking and risk for oral cancer." Cancer Research **56**(22): 5192-5197.

Myers, S., L. Bottolo, et al. (2005). "A fine-scale map of recombination rates and hotspots across the human genome." Science (New York, N.Y.) **310**: 321-324.

Nair, R. P., K. C. Duffin, et al. (2009). "Genome-wide scan reveals association of psoriasis with IL-23 and NF-kappaB pathways." Nature Genetics **41**(2): 199-204.

Nair, R. P., P. E. Stuart, et al. (2006). "Sequence and haplotype analysis supports HLA-C as the psoriasis susceptibility 1 gene." American Journal of Human Genetics **78**(5): 827-851.

Naugler, W. E., T. Sakurai, et al. (2007). "Gender disparity in liver cancer due to sex differences in MyD88-dependent IL-6 production." Science **317**(5834): 121-124.

Naumann, A., J. M. Hempel, et al. (2001). "[Detection of humoral immune response to inner ear proteins in patients with sensorineural hearing loss]." Laryngo- rhino- otologie **80**(5): 237-244.

Neale, B. M., S. E. Medland, et al. (2010). "Meta-analysis of genome-wide association studies of attention-deficit/hyperactivity disorder." Journal of the American Academy of Child and Adolescent Psychiatry **49**(9): 884-897.

Neale, B. M. and P. C. Sham (2004). "The future of association studies: gene-based analysis and replication." American Journal of Human Genetics **75**(3): 353-362.

Nelson, J. L. and M. Ostensen (1997). "Pregnancy and rheumatoid arthritis." Rheumatic diseases clinics of North America **23**(1): 195-212.

Nichols, W. C., N. Pankratz, et al. (2005). "Genetic screening for a single common LRRK2 mutation in familial Parkinson's disease." Lancet **365**(9457): 410-412.

Novembre, J., T. Johnson, et al. (2008). "Genes mirror geography within Europe." Nature **456**(7218): 98-101.

Ober, C., D. a. Loisel, et al. (2008). "Sex-specific genetic architecture of human disease." Nature Reviews Genetics **9**: 911-922.

Ogura, Y., D. K. Bonen, et al. (2001). "A frameshift mutation in NOD2 associated with

susceptibility to Crohn's disease." Nature **411**(6837): 603-606.

Okada, Y., C. Terao, et al. (2012). "Meta-analysis identifies nine new loci associated with rheumatoid arthritis in the Japanese population." Nature Genetics **44**(5): 511-516.

Oksenberg, J. R. and S. E. Baranzini (2010). "Multiple sclerosis genetics--is the glass half full, or half empty?" Nature Reviews Neurology **6**: 429-437.

Orozco, G., J. C. Barrett, et al. (2010). "Synthetic associations in the context of genome-wide association scan signals." Human Molecular Genetics **19**(R2): R137-144.

Pagani, M. R., L. E. Gonzalez, et al. (2011). "Autoimmunity in amyotrophic lateral sclerosis: past and present." Neurology research international **2011**: 497080.

Patsopoulos, N. A., A. Tatsioni, et al. (2007). "Claims of sex differences: an empirical assessment in genetic associations." JAMA : the journal of the American Medical Association **298**(8): 880-893.

Patterson, N., A. L. Price, et al. (2006). "Population structure and eigenanalysis." PLoS Genetics **2**(12): e190.

Pearlman, D. S. (1999). "Pathophysiology of the inflammatory response." The Journal of allergy and clinical immunology **104**(4 Pt 1): S132-137.

Petersen, G. M., L. Amundadottir, et al. (2010). "A genome-wide association study identifies pancreatic cancer susceptibility loci on chromosomes 13q22.1, 1q32.1 and 5p15.33." Nature Genetics **42**(3): 224-228.

Pike, C. J., J. C. Carroll, et al. (2009). "Protective actions of sex steroid hormones in Alzheimer's disease." Frontiers in neuroendocrinology **30**(2): 239-258.

Pohanka, M. (2013). "Impact of melatonin on immunity: a review." Central European Journal of Medicine **8**(4): 369-376.

Price, A. L., G. V. Kryukov, et al. (2010). "Pooled association tests for rare variants in exon-resequencing studies." American Journal of Human Genetics **86**(6): 832-838.

Price, A. L., N. J. Patterson, et al. (2006). "Principal components analysis corrects for stratification in genome-wide association studies." Nature Genetics **38**(8): 904-909.

Pritchard, J. K. (2001). "Are rare variants responsible for susceptibility to complex diseases?" American Journal of Human Genetics **69**(1): 124-137.

Pritchard, J. K. and N. J. Cox (2002). "The allelic architecture of human disease genes: common disease-common variant...or not?" Human Molecular Genetics **11**: 2417-2423.

Purcell, S., B. Neale, et al. (2007). "PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses." American Journal of Human Genetics **81**: 559-575.

Qi, L., M. C. Cornelis, et al. (2010). "Genetic variants at 2q24 are associated with susceptibility to type 2 diabetes." Human Molecular Genetics **19**(13): 2706-2715.

Quintero, O. L., M. J. Amador-Patarroyo, et al. (2012). "Autoimmune disease and gender: plausible mechanisms for the female predominance of autoimmunity." Journal of Autoimmunity **38**(2-3): J109-119.

R Core Team (2013). "R: A Language and Environment for Statistical Computing."

Ramachandran, S., O. Deshpande, et al. (2005). "Support from the relationship of genetic and geographic distance in human populations for a serial founder effect originating in Africa." Proceedings of the National Academy of Sciences, USA **102**(44): 15942-15947.

Randall, J. C., T. W. Winkler, et al. (2013). "Sex-stratified genome-wide association studies including 270,000 individuals show sexual dimorphism in genetic loci for anthropometric traits." PLoS Genetics **9**(6): e1003500.

Rauch, S. D., J. E. San Martin, et al. (1995). "Serum antibodies against heat shock protein 70 in

Meniere's disease." The American journal of otology **16**(5): 648-652.

Reich, D. E., M. Cargill, et al. (2001). "Linkage disequilibrium in the human genome." Nature **411**: 199-204.

Reich, D. E. and E. S. Lander (2001). "On the allelic spectrum of human disease." Trends in genetics : TIG **17**: 502-510.

Remmers, E. F., F. Cosan, et al. (2010). "Genome-wide association study identifies variants in the MHC class I, IL10, and IL23R-IL12RB2 regions associated with Behcet's disease." Nature Genetics **42**(8): 698-702.

Romeo, S., L. A. Pennacchio, et al. (2007). "Population-based resequencing of ANGPTL4 uncovers variations that reduce triglycerides and increase HDL." Nature Genetics **39**(4): 513-516.

Roozendaal, B., B. S. McEwen, et al. (2009). "Stress, memory and the amygdala." Nature Reviews Neuroscience **10**(6): 423-433.

Ropers, H. H. and B. C. Hamel (2005). "X-linked mental retardation." Nature reviews. Genetics **6**(1): 46-57.

Rosenberg, N. A., L. Huang, et al. (2010). "Genome-wide association studies in diverse populations." Nature Reviews Genetics **11**: 356-366.

Ross, M. T., D. V. Grafham, et al. (2005). "The DNA sequence of the human X chromosome." Nature **434**(7031): 325-337.

Routsias, J. G. and A. G. Tzioufas (2006). "The role of chaperone proteins in autoimmunity." Annals of the New York Academy of Sciences **1088**: 52-64.

Rubino, S. J., T. Selvanantham, et al. (2012). "Nod-like receptors in the control of intestinal inflammation." Current Opinion in Immunology **24**(4): 398-404.

Sabatti, C., S. K. Service, et al. (2009). "Genome-wide association analysis of metabolic traits in a birth cohort from a founder population." Nature Genetics **41**(1): 35-46.

Saifi, G. M. and H. S. Chandra (1999). "An apparent excess of sex- and reproduction-related genes on the human X chromosome." Proceedings. Biological sciences / The Royal Society **266**(1415): 203-209.

Sardu, C., E. Cocco, et al. (2012). "Population based study of 12 autoimmune diseases in Sardinia, Italy: prevalence and comorbidity." PloS one **7**(3): e32487.

Sawalha, A. H., R. Webb, et al. (2008). "Common variants within MECP2 confer risk of systemic lupus erythematosus." PloS one **3**(3): e1727.

Sawcer, S., G. Hellenthal, et al. (2011). "Genetic risk and a primary role for cell-mediated immune mechanisms in multiple sclerosis." Nature **476**(7359): 214-219.

Saxena, R., B. F. Voight, et al. (2007). "Genome-wide association analysis identifies loci for type 2 diabetes and triglyceride levels." Science **316**(5829): 1331-1336.

Schaub, M. A., I. M. Kaplow, et al. (2009). "A Classifier-based approach to identify genetic similarities between diseases." Bioinformatics **25**(12): i21-29.

Schizophrenia Psychiatric Genome-Wide Association Study (GWAS) Consortium (2011). "Genome-wide association study identifies five new schizophrenia loci." Nature Genetics **43**(10): 969-976.

Schuster, N. and K. Krieglstein (2002). "Mechanisms of TGF-beta-mediated apoptosis." Cell and Tissue Research **307**(1): 1-14.

Schuurs, A. H. and H. A. Verheul (1990). "Effects of gender and sex steroids on the immune response." Journal of steroid biochemistry **35**(2): 157-172.

Scott, L. J., K. L. Mohlke, et al. (2007). "A genome-wide association study of type 2 diabetes in

Finns detects multiple susceptibility variants." Science **316**(5829): 1341-1345.

Scuteri, A., S. Sanna, et al. (2007). "Genome-wide association scan shows genetic variants in the FTO gene are associated with obesity-related traits." PLoS Genetics **3**(7): e115.

Selmi, C., E. Brunetta, et al. (2012). "The X chromosome and the sex ratio of autoimmunity." Autoimmunity Reviews **11**(6-7): A531-537.

Sen, S., R. M. Nesse, et al. (2003). "A BDNF coding variant is associated with the NEO personality inventory domain neuroticism, a risk factor for depression." Neuropsychopharmacology : official publication of the American College of Neuropsychopharmacology **28**(2): 397-401.

Shatunov, A., K. Mok, et al. (2010). "Chromosome 9p21 in sporadic amyotrophic lateral sclerosis in the UK and seven other countries: a genome-wide association study." Lancet neurology **9**(10): 986-994.

Shen, N., Q. Fu, et al. (2010). "Sex-specific association of X-linked Toll-like receptor 7 (TLR7) with male systemic lupus erythematosus." Proceedings of the National Academy of Sciences of the United States of America **107**(36): 15838-15843.

Sirota, M., M. a. Schaub, et al. (2009). "Autoimmune disease classification by inverse association with SNP alleles." PLoS genetics **5**: e1000792.

Sivakumaran, S., F. Agakov, et al. (2011). "Abundant pleiotropy in human complex diseases and traits." American Journal of Human Genetics **89**(5): 607-618.

Slominski, A., R. Paus, et al. (1989). "Hypothesis - Possible Role for the Melatonin Receptor in Vitiligo - Discussion Paper." Journal of the Royal Society of Medicine **82**(9): 539-541.

Sofaer, J. (1993). "Crohn's disease: the genetic contribution." Gut **34**(7): 869-871.

Solovieff, N., C. Cotsapas, et al. (2013). "Pleiotropy in complex traits: challenges and

strategies." Nature reviews. Genetics **14**(7): 483-495.

Somers, E. C., S. L. Thomas, et al. (2009). "Are individuals with an autoimmune disease at higher risk of a second autoimmune disorder?" American Journal of Epidemiology **169**(6): 749-755.

Sospedra, M. and R. Martin (2005). "Immunology of multiple sclerosis." Annual Review of Immunology **23**: 683-747.

Sowers, J. R. (1998). "Comorbidity of hypertension and diabetes: the fosinopril versus amlodipine cardiovascular events trial (FACET)." The American journal of cardiology **82**(9B): 15R-19R.

Spencer, C., E. Hechter, et al. (2011). "Quantifying the underestimation of relative risks from genome-wide association studies." PLoS genetics **7**(3): e1001337.

Spencer, C. C. a., Z. Su, et al. (2009). "Designing genome-wide association studies: sample size, power, imputation, and the choice of genotyping chip." PLoS genetics **5**: e1000477.

Staal, F. J., T. C. Luis, et al. (2008). "WNT signalling in the immune system: WNT is spreading its wings." Nature reviews. Immunology **8**(8): 581-593.

Stahl, E. A., S. Raychaudhuri, et al. (2010). "Genome-wide association study meta-analysis identifies seven new rheumatoid arthritis risk loci." Nature Genetics **42**(6): 508-514.

Stouffer, S. A., E. A. Suchman, et al. (1949). Adjustment During Army Life. Princeton, NJ, Princeton University Press.

Su, A. I., T. Wiltshire, et al. (2004). "A gene atlas of the mouse and human protein-encoding transcriptomes." Proceedings of the National Academy of Sciences of the United States of America **101**(16): 6062-6067.

Suarez, B. K., J. Duan, et al. (2006). "Genomewide linkage scan of 409 European-ancestry and

African American families with schizophrenia: suggestive evidence of linkage at 8p23.3-p21.2 and 11p13.1-q14.1 in the combined sample." American Journal of Human Genetics **78**(2): 315-333.

Subramanian, A., P. Tamayo, et al. (2005). "Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles." Proceedings of the National Academy of Sciences of the United States of America **102**(43): 15545-15550.

Takeuchi, F., S. Kobayashi, et al. (2011). "Detection of common single nucleotide polymorphisms synthesizing quantitative trait association of rarer causal variants." Genome Research **21**(7): 1122-1130.

Tang, Q. Z. and J. A. Bluestone (2008). "The Foxp3(+) regulatory T cell: a jack of all trades, master of regulation." Nature Immunology **9**(3): 239-244.

Tarpey, P. S., R. Smith, et al. (2009). "A systematic, large-scale resequencing screen of X-chromosome coding exons in mental retardation." Nature Genetics **41**(5): 535-543.

Terry, P. D., F. Villinger, et al. (2009). "Melatonin and Ulcerative Colitis: Evidence, Biological Mechanisms, and Future Research." Inflammatory bowel diseases **15**(1): 134-140.

Teslovich, T. M., K. Musunuru, et al. (2010). "Biological, clinical and population relevance of 95 loci for blood lipids." Nature **466**(7307): 707-713.

The Wellcome Trust Case Control Consortium (2007). "Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls." Nature **447**: 661-678.

Thomas, P. D., M. J. Campbell, et al. (2003). "PANTHER: a library of protein families and subfamilies indexed by function." Genome Research **13**(9): 2129-2141.

Thornton, T., Q. Zhang, et al. (2012). "XM: Association Testing on the X-Chromosome in Case-

Control Samples With Related Individuals." Genetic Epidemiology.

Thurnher, M., H. Clausen, et al. (1992). "T cell clones with normal or defective O-galactosylation from a patient with permanent mixed-field polyagglutinability." European Journal of Immunology **22**(7): 1835-1842.

Tiniakou, E., K. H. Costenbader, et al. (2013). "Sex-specific environmental influences on the development of autoimmune diseases." Clinical Immunology **149**(2): 182-191.

Tishkoff, S., E. Dietzsch, et al. (1996). "Global patterns of linkage disequilibrium at the CD4 locus and modern human origins." Science.

Todd, J. A., N. M. Walker, et al. (2007). "Robust associations of four new chromosome regions from genome-wide analyses of type 1 diabetes." Nature Genetics **39**(7): 857-864.

Troutbeck, R., S. Al-Qureshi, et al. (2012). "Therapeutic targeting of the complement system in age-related macular degeneration: a review." Clinical & experimental ophthalmology **40**(1): 18-26.

Tukiainen, T., M. Pirinen, et al. (2014). "Chromosome x-wide association study identifies Loci for fasting insulin and height and evidence for incomplete dosage compensation." PLoS Genetics **10**(2): e1004127.

Tysk, C., E. Lindberg, et al. (1988). "Ulcerative colitis and Crohn's disease in an unselected population of monozygotic and dizygotic twins. A study of heritability and the influence of smoking." Gut **29**(7): 990-996.

Ventriglia, M., L. Bocchio Chiavetto, et al. (2002). "Association between the BDNF 196 A/G polymorphism and sporadic Alzheimer's disease." Molecular Psychiatry **7**(2): 136-137.

Wang, J., D. Duncan, et al. (2013). "WEB-based GEne SeT AnaLysis Toolkit (WebGestalt): update 2013." Nucleic Acids Research.

Wang, K., S. P. Dickson, et al. (2010). "Interpretation of association signals and identification of causal variants from genome-wide association studies." American Journal of Human Genetics **86**(5): 730-742.

Warde-Farley, D., S. L. Donaldson, et al. (2010). "The GeneMANIA prediction server: biological network integration for gene prioritization and predicting gene function." Nucleic Acids Research **38**(Web Server issue): W214-220.

Weatherhead, S. C., P. M. Farr, et al. (2011). "Keratinocyte apoptosis in epidermal remodeling and clearance of psoriasis induced by UV radiation." The Journal of investigative dermatology **131**(9): 1916-1926.

Weishaupt, J. H., C. Bartels, et al. (2006). "Reduced oxidative damage in ALS by high-dose enteral melatonin treatment." Journal of Pineal Research **41**(4): 313-323.

Welter, D., J. MacArthur, et al. (2014). "The NHGRI GWAS Catalog, a curated resource of SNP-trait associations." Nucleic Acids Research **42**(Database issue): D1001-1006.

Whirl-Carrillo, M., E. M. McDonagh, et al. (2012). "Pharmacogenomics knowledge for personalized medicine." Clinical pharmacology and therapeutics **92**(4): 414-417.

Whitacre, C. C. (2001). "Sex differences in autoimmune disease." Nature Immunology **2**(9): 777-780.

Whitacre, C. C., S. C. Reingold, et al. (1999). "A gender gap in autoimmunity." Science **283**(5406): 1277-1278.

Willer, C. J., Y. Li, et al. (2010). "METAL: fast and efficient meta-analysis of genomewide association scans." Bioinformatics **26**(17): 2190-2191.

Wilson Sayres, M. A. and K. D. Makova (2013). "Gene survival and death on the human Y chromosome." Molecular Biology and Evolution **30**(4): 781-787.

Winfield, J. B. and W. N. Jarjour (1991). "Stress proteins, autoimmunity, and autoimmune disease." Current Topics in Microbiology and Immunology **167**: 161-189.

Wise, A. L., L. Gyi, et al. (2013). "eXclusion: toward integrating the X chromosome in genome-wide association analyses." American Journal of Human Genetics **92**(5): 643-647.

Wooten, G. F., L. J. Currie, et al. (2004). "Are men at greater risk for Parkinson's disease than women?" Journal of neurology, neurosurgery, and psychiatry **75**(4): 637-639.

Wray, N. R., S. M. Purcell, et al. (2011). "Synthetic associations created by rare variants do not explain most GWAS results." PLoS Biology **9**(1): e1000579.

Wu, C., I. Macleod, et al. (2013). "BioGPS and MyGene.info: organizing online, gene-centric information." Nucleic Acids Research **41**(Database issue): D561-565.

Wu, M. C., S. Lee, et al. (2011). "Rare-variant association testing for sequencing data with the sequence kernel association test." American Journal of Human Genetics **89**(1): 82-93.

Yancik, R., R. J. Havlik, et al. (1996). "Cancer and comorbidity in older patients: a descriptive profile." Annals of Epidemiology **6**(5): 399-412.

Yang, J., T. a. Manolio, et al. (2011). "Genome partitioning of genetic variation for complex traits using common SNPs." Nature Genetics.

Zaccara, G. (2009). "Neurological comorbidity and epilepsy: implications for treatment." Acta neurologica Scandinavica **120**(1): 1-15.

Zang, E. A. and E. L. Wynder (1996). "Differences in lung cancer risk between men and women: examination of the evidence." Journal of the National Cancer Institute **88**(3-4): 183-192.

Zaykin, D. V., L. A. Zhivotovsky, et al. (2002). "Truncated product method for combining P-values." Genetic Epidemiology **22**(2): 170-185.

Zhang, B., S. Kirov, et al. (2005). "WebGestalt: an integrated system for exploring gene sets in

various biological contexts." Nucleic Acids Research **33**(Web Server issue): W741-748.

Zheng, G., J. Joo, et al. (2007). "Testing association for markers on the X chromosome." Genetic Epidemiology **31**(8): 834-843.

Zhernakova, A., E. A. Stahl, et al. (2011). "Meta-analysis of genome-wide association studies in celiac disease and rheumatoid arthritis identifies fourteen non-HLA shared loci." PLoS Genet **7**(2): e1002004.