bioRxiv preprint doi: https://doi.org/10.1101/009464; this version posted September 21, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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

Diana Chang[1,2], Feng Gao[1], Li Ma[1,3], Aaron J. Sams[1], Andrea Slavney[1,4], Yedael Y. Waldman[1], Paul Billing-Ross[1,4], Aviv Madar[1], Richard Spritz[5], Alon Keinan[1,2]

Running Title: X-linked genes implicated in autoimmune diseases

[1] Department of Biological Statistics & Computational Biology, Cornell University, Ithaca, New York, United States of America
[2] Program in Computational Biology and Medicine, Cornell University, Ithaca, New York, United States of America
[3] Department of Animal and Avian Sciences, University of Maryland, College Park, MD, United States of America
[4] Genetics, Genomics & Development, Cornell University, Ithaca, New York, United States of America
[5] Human Medical Genetics and Genomics Program, University of Colorado School of Medicine, Aurora, CO, United States of America

bioRxiv preprint doi: https://doi.org/10.1101/009464; this version posted September 21, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

ABSTRACT

Many complex human diseases are highly sexually dimorphic, which suggests a potential contribution of the X chromosome. However, the X chromosome has been neglected in most genome-wide association studies (GWAS). We present tailored analytical methods and software that facilitate X-wide association studies (XWAS), which we further applied to reanalyze data from 16 GWAS of different autoimmune diseases (AID). We associated several X-linked genes with disease risk, among which *ARHGEF6* is associated with Crohn's disease and replicated in a study of ulcerative colitis, another inflammatory bowel disease (IBD). Indeed, ARHGEF6 interacts with a gastric bacterium that has been implicated in IBD. Additionally, we found that the centromere protein CENPI is associated with three different AID; replicated a previously investigated association of *FOXP3,* which regulates genes involved in T-cell function, in vitiligo; and discovered that *C1GALT1C1* exhibits sex-specific effect on disease risk in both IBDs. These and other X-linked genes that we associated with AID tend to be highly expressed in tissues related to immune response, display differential gene expression between males and females, and participate in major immune pathways. Combined, the results demonstrate the importance of the X chromosome in autoimmunity, reveal the potential of XWAS, even based on existing data, and provide the tools and incentive to appropriately include the X chromosome in future studies.

bioRxiv preprint doi: https://doi.org/10.1101/009464; this version posted September 21, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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 [1]. As a consequence, though X constitutes 5% of the nuclear genome and underlies almost 10% of Mendelian disorders [2-4], it harbors only 15 out of the 2,800 (0.5%) total significant associations for nearly 300 traits [1,5,6]. This 0.5% of associated SNPs is less often in putatively functional loci compared to autosomal associated SNPs [1,5,7], which further suggests that X-linked associations might contain a higher proportion of false positives. This is likely due to most studies analyzing X with tools that were designed for the autosomes [1]. We hypothesize that X explains a portion of "missing heritability" [8,9], especially for the many complex human diseases that exhibit gender disparity in risk, age of onset, or symptoms. In fact, the complex human diseases most extensively studied in GWAS are highly sexually dimorphic, including autoimmune diseases [10-12], neurological and psychiatric disorders [13-17], cardiovascular disease [18-22], and cancer [23-26]. Several mechanisms underlying sexual dimorphism have been suggested [12,27-31], including the contribution of the X chromosome [27,32-35]. The hypothesis is further motivated by the importance of X in sexually dimorphic traits in both model organisms and human Mendelian disorders, as well as by its enrichment for sexually antagonistic alleles, which are expected to disproportionately contribute to complex disease risk [36]. Finally, 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 sets of genes with unique functions [37-39].

bioRxiv preprint doi: https://doi.org/10.1101/009464; this version posted September 21, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

X-specific problems that should be taken into consideration in GWAS 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 [40], 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 [41-47]; 4) quality control (QC) criteria that account for sex information to prevent filtering the entirety or a large fraction of the chromosome [1], while at the same time accounting for confounding sex-specific effects; 5) sex-specific population structure leading to differential effects of population stratification (which could inflate the type I error rate [48-50]) between X and the autosomes; and 6) application of association tests designed for the autosomes, which leads to statistical inaccuracy. Recent advancements of association test statistics for X have been made [51-57], with a recent important study discovering X-linked loci associated to height and fasting insulin [56].

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) [10-12,58]. Furthermore, they often show sex-specific symptoms, age of onset, and progression [10-12,29,59-62]. While pregnancy [12,30,31] and other environmental factors [63], as well as sex hormones [12,29-31], can contribute to sexually dimorphic characteristics, a role for X-linked genes has also been suggested [27,62,64-66]. Though AID have been extensively

bioRxiv preprint doi: https://doi.org/10.1101/009464; this version posted September 21, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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 [67-69]. Across the dozens of GWAS in AID, few have studied the contribution of X and, to date, few have provided evidence of its role in AID [1,5,6].

In this study, we first introduce X-specific analytical methods and software for carrying out XWAS, which take into account several of the aforementioned problems. They apply X-specific strategies for QC, imputation, association methods, and tests of sex-specific effects. Furthermore, motivated by the unique nature of genes on X, we implemented and applied the first gene-based tests for associating X-linked genes. We apply these methods to conduct an extensive XWAS of a number of AID and other diseases with a potential autoimmune component (DPACs) [70,71]. Our discovery of X-linked risk genes illuminate the potential importance of X in autoimmune disease etiology, 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.

bioRxiv preprint doi: https://doi.org/10.1101/009464; this version posted September 21, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

RESULTS AND DISCUSSION

Associations of X-linked genes with autoimmune disease risk

We assembled 16 datasets of AID and DPACs for analysis (Table 1). To facilitate independent analysis and replication in these datasets, we removed some individuals such that no overlapping data remains between the 16 datasets (Materials and Methods). For each dataset, we first carried out QC that was 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) carried out a different analysis for X [72]. We applied three statistical methods to measure disease association for each SNP in each of the datasets (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). Second, we employed regression analyses separately for each sex and combined them into a single test of association using either 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. We employed EIGENSOFT [48] to remove individuals of non-European descent and correct for potential population stratification. Following this correction, QQ (quantile-quantile) plots of the three tests across all SNPs, along with genomic inflation factors for each dataset, revealed no systematic bias (Figure S1; Table S1).

bioRxiv preprint doi: https://doi.org/10.1101/009464; this version posted September 21, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Based on the three test statistics, we conducted gene-based tests that each combines all tests of individual SNPs that span a gene (Materials and Methods) [73-76]. As hemizygosity in males and X-inactivation reduces the effective sample size for X, this test provides a crucial improvement in statistical power by focusing on whole genes as functional units and reducing the multiple hypothesis testing burden from the number of SNPs to the number of genes [74,75]. This approach also surmounts issues of replication across studies with different sets of SNPs that arise from differing genotyping arrays and quality control filtering. We tested each gene in each of the 16 datasets using each of FM_{02}, $FM_{F.comb}$ and $FM_{S.comb}$ tests as described in the following. For completeness, results for individual SNP tests are provided in Supplementary Text, Figure S2, and Tables S2-S3.

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. The gene-based test for aggregates signals across all SNPs in each gene, while accounting for the structure of linkage disequilibrium (LD) within each gene (Materials and Methods). Combining signals across all SNPs was based on truncated tail strength [77] and truncated product [78] methods. Rather than consider only the single SNP with the strongest signal, these methods combine signals from several most significant SNPs, thus improving statistical power, especially in cases where a gene contains multiple risk alleles or where the causal SNP is partially tagged by different tested SNPs (Materials and Methods) [79,80]. From the first round of discovery, we considered for replication genes with significance of $P < 10^{-3}$ (Tables S4-S5). For these, we first attempted replication in a different dataset of the same disease (including the related Crohn's

bioRxiv preprint doi: https://doi.org/10.1101/009464; this version posted September 21, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

disease and ulcerative colitis), if such a dataset was available for our analysis (Table 1), while applying Bonferroni correction for the number of genes we attempted to replicate (Table 1). Otherwise, motivated by the shared pathogenicity of different AID [81-84] (which is also supported by our following results), we attempted replication in all other datasets considered herein (Table 1). In both cases, we considered replication using the same test statistic that passed the discovery significance criterion.

We detected 54 unique genes that passed the initial criteria for discovery in one or more of the 16 datasets. Of these, 38 genes were significant based on the FM_{02} test, 22 based on the $FM_{F.comb}$ test, and 34 in the $FM_{S.comb}$ test (Tables S4-S5), with overlap between the three tests due to their statistical dependence. For 42 genes out of the 54, we had an independent data set of the same or related disease with which to attempt replication. Out of these 42, 5 successfully replicated, with 3 of the 5 both discovered and replicated based on more than one of the three tests (Figure 1a-c and Table 2). 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. 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 remain in LD with the same—likely unobserved—causal variant(s).

FOXP3, which we associated with vitiligo risk (combined $P = 9.5x10^{-6}$; Table 2) has been previously associated to vitiligo in candidate gene study [85]. Vitiligo is a common autoimmune disorder that is manifested in patches of depigmented skin due to abnormal destruction of

bioRxiv preprint doi: https://doi.org/10.1101/009464; this version posted September 21, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

melanocytes. *FOXP3* 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 [86,87]. Defects in the gene are also a known cause for an X-linked Mendelian autoimmunity-immunodeficiency syndrome (IPEX - immunodysregulation polyendocrinopathy enteropathy X-linked syndrome) [88].

In Crohn's Disease (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 it in the Wellcome Trust Case Control Consortium 2 (WT2) ulcerative colitis, another inflammatory bowel disease (combined $P = 1.67 \times 10^{-5}$). ARHGEF6 binds to a major surface protein of *H. pylori* [89], which is a gastric bacterium that may play a role in inflammatory bowel diseases [90,91].

We discovered that another gene, *CENPI*, was associated with three diseases (celiac disease, vitiligo, and amyotrophic lateral sclerosis (ALS)), with a combined $P = 2.1 \times 10^{-7}$ (Table S6). 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 ($P = 2.7 \times 10^{-5}$). *CENPI* is a member of a protein complex that generates spindle assembly checkpoint signals required for cell progression through mitosis [92]. A previous study demonstrated that it is targeted by the immune system in some scleroderma patients [93]. Additionally, autosomal genes in the same family have been previously associated with immune-related diseases, such as multiple sclerosis (*CENPC1*) [94] and ALS (*CENPV*) [95]. These findings combined suggest a possible general role for *CENPI* in autoimmunity.

bioRxiv preprint doi: https://doi.org/10.1101/009464; this version posted September 21, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Motivated by the association of *CENPI* in multiple AID and DPACs, as well as previous evidence of shared pathogenicity across different AID [81,82], we next sought to replicate the 54 genes from the discovery stage in diseases different from those in which they were discovered (using the same test statistic as in their discovery). We successfully replicated 17 genes in this fashion, on top of the aforementioned 5 that replicated in the same or related disease (Figure 1a-c and Table 3). Six of the 17 were both discovered and replicated based on more than one of the three test statistics, and 5 of the 17 replicated in two separate datasets. We consider these results based on replication in other diseases to provide only suggestive evidence of these genes playing a role in autoimmunity or immune-response, and consider these genes together with the above 5 in subsequent analyses.

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

If X-linked genes underlie 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). No evidence for systemic bias was observed (Figure S3; Table S1). As with our above analyses, we combined SNP-level results to a gene-based test of sex-differentiated effect size. This test captures a scenario whereby SNPs within the tested gene display different effects in males and females, without constraining such differential effects to be of a similar nature across SNPs. Detailed results are provided in Figure 1d, Tables 2-3, and Table S7. Specifically, we discovered and replicated *C1GALT1C1* as exhibiting sex-differentiated effect size in risk of IBD (combined P = 4.11×10^{-5}). C1GALT1C1 (also known as Cosmc) is necessary for the synthesis of many O-

bioRxiv preprint doi: https://doi.org/10.1101/009464; this version posted September 21, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

glycan proteins [96], which are components of several antigens. Defects of C1GALT1C1 may cause Tn Syndrome (a hematological disorder) [97]. We also considered replication of sex-differentiated effect in diseases different from those in which a sex-differentiated effect had been discovered. This replication found 8 additional genes, including both *CENPI* (combined P = 1.6×10^{-8}) and *MCF2* (combined P = 2.0×10^{-4}), which we associated to risk of AID in our analyses above. The evidence of the sex-differentiated effect of these genes is in the same diseases as in the association analysis, thereby pointing to a sex-specific effect on disease risk in these diseases (Figure 1d and Table 3). We stress again replication in other diseases is only to be considered as suggestive evidence, and consider for subsequent analyses these genes together with the genes replicated in the same disease.

Sex-differentiated effects can be a consequence of the X-inactivation status of the gene. X-inactivation is a dosage compensation mechanism between XX females and XY males that silences transcription from one of the two X chromosomes in each female somatic cell. However, at least 25% of human X loci escape X-inactivation to varying degrees. Studying the literature for the X-inactivation status of the 3 above genes (*C1GALT1C1*, *CENPI*, and *MCF2*) we found no evidence that they do not undergo complete inactivation [33,98]. Furthermore, all three have degenerate Y gametologs in males, i.e. either the gene has been lost from the Y chromosome (*MCF2*) or their homologous gene on the Y is a nonfunctional pseudogene (*C1GALT1C1* and *CENPI)*. (In the following section, we specifically test the set of X-linked genes with functional Y homologs). Based on these observations, these genes are not expected to show sex-differential expression, at least in fibroblasts, in which X-inactivation statuses has been derived [33,98]. However, it is possible that these genes show female-biased expression in other tissues as a

bioRxiv preprint doi: https://doi.org/10.1101/009464; this version posted September 21, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

consequence of them escaping X-inactivation in a tissue-specific or disease-specific manner [99,100]. Hence, in the following we more directly test for sex-differential gene expression. Additionally, the sex-differential risk factor may arise from interaction with other genes and sex-specific environmental factors.

As X-inactivation status does not directly support differential expression between the sexes, we next tested whether any of the X-linked genes we associated with AID exhibit differences in expression between males and females. We considered a comprehensive dataset of whole blood gene expression from 881 individuals (409 males and 472 females; Materials and Methods) and assayed gene expression in males and females separately. Considering all X-linked genes that we analyzed, they exhibit a 2.55-fold enrichment for differential expression between males and females as compared to all genes across all chromosomes ($P=6.5x10^{-8}$). *XIST*, the gene responsible for X-inactivation, displays the most significant difference between males and females among all X-linked genes ($P<<10^{-16}$). Within the genes we associated with AID risk, four exhibit 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, two of these genes (*EFHC2* and *BEND2*) also exhibit sex-differentiated effect sizes in the above analysis, though replicated as such in AID different from the ones in which they were discovered (Figure 1d and Table 3). Hence, the results herein propose that sex-differentiated effect sizes are potentially related to sex-differential expression pattern.

Tissue-specific expression of X-linked genes implicated in autoimmune disease risk

bioRxiv preprint doi: https://doi.org/10.1101/009464; this version posted September 21, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

As the X chromosome carries 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 (Figure 2-3): *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* (associated with ALS), and *TMEM35* (associated with ALS) exhibit the highest expression levels in the pineal gland. The pineal gland produces and secretes melatonin, which interacts with the immune system [101,102] and has been implicated in both vitiligo and ALS [101,103-107], as well as suggested as a possible treatment for ALS [108]. 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 [109,110], which is believed to play a significant role in susceptibility to psoriasis [111].

Association of genes with Y homologs and immune-related function

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 AID disease risk. We tested this hypothesis independent of the above results by testing for a concurrent association of a whole gene set suggested by each hypothesis with each of the diseases (Materials and Methods). We considered 3 such gene sets, with the first two sets including 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

bioRxiv preprint doi: https://doi.org/10.1101/009464; this version posted September 21, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

analysis of the immune-related gene sets has been motivated by the nature of the diseases, the test of the latter set has been motivated by an evolutionary perspective. These genes are more likely to be under functional constraint since their Y homologs have retained function despite loss off recombination with X (which has led to progressive degeneration of the Y chromosome over the course of the evolution of the supercohort *Theria*) [98]. Thus, they may be more likely to play a part in disease etiology.

The Panther immunity gene set is associated with vitiligo risk in both vitiligo studies that we analyzed and in 3 test statistics of association, as well associated with one type-2 diabetes study based on the $FM_{S.comb}$ test statistic (Table 4). Similarly, the KEGG/GO set is associated in one of the vitiligo datasets GWAS1 (Table 4). The set of genes with functional Y homologs suggestively contribute to a much larger set of diseases, including psoriasis, vitiligo, celiac disease, Crohn's Disease, and type 1 diabetes, with the first two of these significant following Bonferroni correction (Table 4). See Table S8 for detailed results for all other datasets and tests.

Relation between associated disease risk genes

We next investigated the relationship between the combined set of X-linked genes we discovered and replicated as associated with any AID or DPAC. First, we considered co-expression of these genes across 881 individuals (Materials and Methods). We observed that 3.9% of all X-linked gene pairs exhibit significant positively correlated gene expression patterns. Pairs of genes from the combined set associated with any AID or DPAC exhibit significantly-positively correlated expression for 8% of pairs - a significantly higher fraction relative to X-linked genes overall

bioRxiv preprint doi: https://doi.org/10.1101/009464; this version posted September 21, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

(Table S9; $P=1.53x10^{-3}$). This suggests that these genes are more likely to work in concert and perhaps interact in the same pathways or cellular networks.

Motivated by their co-expression, we next investigated the relationship of the combined set of X-linked genes we discovered and replicated using data from protein-protein or genetic interactions (Materials and Methods). We found that all but 4 of the 22 genes are included in the same interaction network (Figure 4). In a pathway enrichment analysis of the resulting interactome (i.e. all genes in Figure 4), several of the significantly enriched pathways relate to immune response or specific immune-related disorders or diseases (Table 5). Other significantly enriched pathways include the regulation of actin cytoskeleton, which has been previously found to influence the developing morphology and movement of T-cells, as well as the TGF-beta signaling and ECF-receptor interaction pathways that can both mediate apoptosis [112,113]. Finally, the significantly enriched Wnt signaling pathway is generally involved in cell development processes, such as cell-fate determination and cell differentiation [114]. 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 [115].

Concluding remarks

In this study, we applied an X-tailored analysis pipeline to 16 different GWAS datasets (Table 1), discovered and replicated several genes associated with autoimmune disease risk (Figure 1, Tables 2-3). Multiple additional lines of evidence point to some of these genes having immune-related functions, including expression in immune-related tissues (Figure 2) and enrichment of these genes in immune-related pathways (Table 5; Figure 4). Beyond immune function, several

bioRxiv preprint doi: https://doi.org/10.1101/009464; this version posted September 21, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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 [116-118] and shows strong evidence for involvement in the etiology of vitiligo [119], psoriasis [120] and rheumatoid arthritis [121]. 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 (Figure 1, Tables 2-3).

The X chromosome has received attention in GWAS during the past year [1,56,122,123]. Our results highlight the contribution of X to autoimmune diseases risk and yields new avenues for potential functional follow-ups, including unraveling the sexual dimorphism of autoimmune diseases. 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 GWAS, we have released our software for handling chromosome X [124] (http://keinanlab.cb.bscb.cornell.edu/content/tools-data), which we provide as an extension of PLINK [55]. Further expansions of this initial software can take advantage of unique features of the X chromosome to further develop X-tailored methods such as methods that rely on X-inactivation and on the availability of phased X haplotypes from males.

bioRxiv preprint doi: https://doi.org/10.1101/009464; this version posted September 21, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Materials and Methods

Datasets

All GWAS datasets used in this study are summarized in Table 1. Out of these, we obtained the following datasets from dbGaP: ALS Finland [125] (phs000344), ALS Irish [126] (phs000127), Celiac disease CIDR [127] (phs000274), MS Case Control [94] (phs000171), Vitiligo GWAS1 [128] (phs000224), CD NIDDK [129] (phs000130), CASP [130] (phs000019), and T2D GENEVA [131] (phs000091).

Additional datasets were obtained from the Wellcome Trust Case Control Consortium (WT): all WT1 [72] datasets, WT2 ankyolosing spondylitis (AS) [132], WT2 ulcerative colitis (UC) [133] and WT2 multiple sclerosis (MS) [134] (Table 1). In order for replication tests and meta-analysis to not be biased by these datasets containing some of the same control samples, we removed all overlap. 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. Finally, 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.

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 [70,71].

bioRxiv preprint doi: https://doi.org/10.1101/009464; this version posted September 21, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Lastly, we analyzed The Vitiligo GWAS2 dataset [135], which contained case data only, as is also the case for the Vitiligo GWAS1 that we downloaded from dbGaP. Therefore, we obtained the following additional datasets from dbGaP: PanScan [136,137] (phs000206), National Institute on Aging Alzheimer's study [138] (phs000168), CIDR bone fragility [139] (phs000138), COGA [140] (phs000125), and SAGE [140-142] (phs000092). Only samples with the "general research consent" designation in these control datasets were used as controls for studying vitiligo. These samples were randomly distributed between the Vitiligo GWAS1 and Vitiligo GWAS2 datasets.

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 [143]. We additionally filtered variants with >10% missingness, variants with a minor allele frequency (MAF) < 0.005, and variants for which missingness was significantly correlated with phenotype ($P<1x10^{-4}$). X-specific QC steps included filtering variants that are not in Hardy-Weinberg equilibrium in females ($P<1x10^{-4}$) or that had significantly different MAF between males and females in control individuals (P<0.05/#SNPs), as well as removal of the pseudoautosomal regions (PARs). We also implemented and considered sex-stratified QC, namely filtering X-linked variants and individuals via separate QC in males and females [122]. However, since we observed no difference in the significant results when applying it to two of the datasets (CD NIDDK, MS case

bioRxiv preprint doi: https://doi.org/10.1101/009464; this version posted September 21, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

control), we considered for analysis data prior this QC step. Finally, following all above QC steps, we removed variants that exhibit differential missingness between males and females [122,144,145]. This step follows the procedure described by König et al. [122] based on a [illegible] test ($P<10^{-7}$).

Correction for population stratification

Sex-biased demographic events, including differential historical population structure of males and females have been proposed for many human populations (e.g. [43,47,146-148]). Such sex-biased history is expected to lead to differential population structure on X and the autosomes, thus to differential population stratification. Essentially, population structure of the X chromosome captures the appropriate mix of male- and female-population structure that is required for an association study of X-linked loci. Hence, we assessed and corrected for potential population stratification via either autosomal-derived or X-derived principal components, and studied the inflation of test statistics in each case as observed in QQ plots. This was performed by principal component analysis (PCA) using EIGENSOFT [48], after pruning for linkage disequilibrium (LD) and removing large LD blocks [50]. For all the datasets analyzed here, which all consist solely of individuals 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. We note, however, that in association studies where more data is available for X, or in other populations—such as population that exhibit a higher level of historical sex-biased population structure—consideration

bioRxiv preprint doi: https://doi.org/10.1101/009464; this version posted September 21, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

of population structure on the X chromosome alone can provide a more accurate population stratification correction for XWAS.

All subsequent analyses are hence based on first excluding any individuals inferred based on EIGENSOFT [48] to be of non-European ancestry. Then, assessment and correction for population stratification follow the convention of using the first ten autosomal-derived PCs as covariates [49], which is supported by investigation of the resultant QQ plots and by population stratification reported by the studies which data we analyzed. Principal component covariates were not added to the regression model only for the amyotrophic lateral sclerosis (ALS) Finland, ALS Irish, and CASP datasets as no inflated p-values were observed in these studies [125,126,130] (Figure S1).

Imputation

Imputation was carried out with IMPUTE2 [149] version 2.2.2 based on 1000 Genomes Project [150] 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 1 displays the number of SNPs we considered in each dataset following imputation and these additional QC steps.

bioRxiv preprint doi: https://doi.org/10.1101/009464; this version posted September 21, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Single-SNP association analysis

We considered 3 tests for associated X-linked SNPs with disease risk. The first test effectively assumes complete and uniform X-inactivation in females and a similar effect size between males and females. In this test, females are hence considered to have 0, 1, or 2 copies of an allele as in an autosomal analysis. Males are considered to have 0 or 2 copies of the same allele, i.e. male hemizygotes are considered equivalent to female homozygous states. This test is implemented in PLINK [55] as the *–xchr-model 2* option, termed FM_{02} in this study. We do note that the assumptions of complete X-inactivation and equal effect sizes often do not hold (see also tests and results of sex-differentiated effect size and sex-differentiation gene expression). Hence, in the second test, termed $FM_{F.comb}$, data from each sex (cases and controls) 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 Fisher's method [151]. This test accommodates the possibility of differential effect size and direction between males and females and is not affected by the allele coding in males (e.g. 0/2 copies or 0/1 copies). Finally, the third test, termed $FM_{S.comb}$, mirrors the second except for using a weighted Stouffer's method [152] instead of Fisher's method. While Fisher's method combines the final p-values, Stouffer's method allows combining and weighing of test statistics. The male-based and female-based test statistics are weighted by the square-root of the male or female sample size [153], and combined while also taking into account the direction of effect in males and females. Equations are implemented as provided in Willer *et al.* [153]. Power calculation of these 3 test statistics for different effect sizes and sample sizes is provided in Supplementary text and Figures S4-S7.

Gene-based association analysis

bioRxiv preprint doi: https://doi.org/10.1101/009464; this version posted September 21, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Based on all single-SNP association tests that use a certain test statistics, we implemented an equivalent gene-based test for each of the 3 test statistics by considering all SNPs across each gene. This was carried out in the general framework of VEGAS [73], a brief description of which follows. VEGAS utilizes the LD between SNPs in a gene to derive the distribution of test statistics [73]. 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. The gene-based p-value is then calculated as the proportion of simulated statistics that were as or more extreme than the observed statistic.

Here, we have implemented a slight modification to this procedure: Instead of a summation or the minimum p-value approach, we combined p-values derived from the simulated test statistics with either the truncated tail strength [77] or the truncated product [78] method, which have been suggested to be more powerful in some scenarios [79,80]. To increase time efficiency of the simulation procedure, adaptive simulations were implemented as in VEGAS [73]. We obtained a list of X-linked genes and their positions from the 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. When several genes in LD show a significant signal, we repeated analysis while removing the flanking 15 kb on each side.

Test of sex-differential effect size

bioRxiv preprint doi: https://doi.org/10.1101/009464; this version posted September 21, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

The difference in the effect size between males and females at each SNP was assayed based on the sex-stratified test described above. Considering the female-only and male-only from that test, differential effect size is tested using the following t-statistic [154]:

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

where *OR* are the odds ratio estimated in either the male-only or female-only tests, *SE* are their respective standard errors, and r the Spearman rank correlation coefficient between $\log(OR_{male})$ and $\log(OR_{female})$ across all X-linked SNPs. For the odds ratios to be comparable, the odds ratio in males is estimated with coding as having 0 or 2 copies. Power calculation of this test statistics for different effect sizes and sample sizes is provided in Supplementary text. Finally, we combined the single-SNP tests in each gene into a gene-based test of sex-differential effect size, along the same lines described above for the association test statistics.

Sex-differential and co-expression gene expression testing

Whole blood gene expression data for 881 samples (409 males, 472 females) from the Rotterdam Study III [155] was downloaded from Gene Expression Omnibus [156] (accession GSE33828). Expression data was available for 803 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 803 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 22 genes that were associated and replicated in any dataset (Figure 1; Tables 2-3) showed significant differential expression between males and

bioRxiv preprint doi: https://doi.org/10.1101/009464; this version posted September 21, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

females. Expression data is available for 20 of these genes, and Bonferroni correction was applied based on 20 tests.

We tested for co-expression between X-linked genes using the non-parametric Spearman's rank correlation test 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 and replicated genes as compared to all 803 genes was tested using a hypergeometric test.

Tissue-specific gene expression

For analysis of tissue-specific gene expression, we obtained the Human GNF1H tissue-specific expression dataset [157] via the BioGPS website [158]. 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 13 of the genes that were associated and replicated in any dataset. 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.

Network analysis

A network of interacting genes was assembled in GeneMANIA using confirmed and predicted genetic and protein interactions [159] with a seed list of the 22 protein-coding genes that were associated and replicated across all datasets (Figure 1; Tables 2-3). To minimize bias towards well-studied pathways, all gene-gene, protein-protein and predicted interaction sub-networks were given equal weight when combined into the final composite network. The resulting composite network consisted of the 22 seed genes and the 100 highest-scoring genetic, protein-

bioRxiv preprint doi: https://doi.org/10.1101/009464; this version posted September 21, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

protein, and predicted interactors. A list of unique genes within this interactome was extracted as input to WebGestalt [160,161] to discover the ten most significantly enriched pathways in the KEGG [162] database. Enrichment was assessed with the hypergeometric test [160] and reported p-values were adjusted for multiple testing using the Benjamini-Hochberg correction as suggested for such analyses [160]. Pathways with a single gene in the interactome were excluded.

Gene set tests

We additionally tested whether SNPs in a pre-compiled set of genes were collectively associated with disease risk. To accomplish this, we modified the gene-based analysis described above to consider multiple genes. Specifically, the simulation step now entails drawing from m multivariate normal distributions, with m denoting the number of genes in the tested gene set. Each of the m multivariate normal distributions denotes one gene and has its own covariance matrix that corresponds to the LD between SNPs in that gene. To verify that this procedure, previously proposed for gene-based tests, can be applied to gene sets, we compared p-values derived from phenotypic permutations to this simulation procedure, which revealed highly correlated significance values (Figures S8-S9). Thus, we only present results the simulation procedure, rather than from a limited number of computationally-intensive permutations.

We applied this test to 3 sets of genes: (1) We manually curated a set of immune-related genes from the KEGG [162] pathways and Gene Ontology (GO) [163] biological function categories. We first these two databases using 15 and 14 categories, respectively, that are particularly relevant for autoimmune response. We subsequently removed eight genes from this list that we

bioRxiv preprint doi: https://doi.org/10.1101/009464; this version posted September 21, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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 (Table S10); (2) The Panther immune gene set was obtained by including all genes in the category of "immune system processes" in the Panther database [164]; and (3) The XY homolog gene set was obtained from data provided by a recent paper by Wilson-Sayres & Makova [98].

bioRxiv preprint doi: https://doi.org/10.1101/009464; this version posted September 21, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

ACKNOWLEDGEMENTS

Some of the datasets used for the analyses described in this manuscript were obtained through dbGaP accession numbers phs000344, phs000127, phs000274, phs000171, phs000224, phs000130, phs000019, phs000091, phs000206, phs000168, phs000138, phs000125 and phs000092. We thank the NIH data repository, the contributing investigators who contributed the phenotype data and DNA samples from their original study, and the primary funding organizations that supported the contributing studies.

This study makes use of data generated by the Wellcome Trust Case Control Consortium. A full list of the investigators who contributed to the generation of the data is available from www.wtccc.org.uk. Funding for the project was provided by the Wellcome Trust under award 076113.

The title of our paper is inspired by the title of a recent review by A. Wise, L. Gyi, and T.A. Manolio that called for inclusion of chromosome X in association studies [1].

bioRxiv preprint doi: https://doi.org/10.1101/009464; this version posted September 21, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

REFERENCES

1. Wise AL, Gyi L, Manolio TA (2013) eXclusion: toward integrating the X chromosome in genome-wide association analyses. Am J Hum Genet 92: 643-647.
2. Hamosh A, Scott AF, Amberger JS, Bocchini CA, McKusick VA (2005) Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. Nucleic Acids Res 33: D514-517.
3. Amberger J, Bocchini CA, Scott AF, Hamosh A (2009) McKusick's Online Mendelian Inheritance in Man (OMIM). Nucleic Acids Res 37: D793-796.
4. Amberger J, Bocchini C, Hamosh A (2011) A new face and new challenges for Online Mendelian Inheritance in Man (OMIM(R)). Hum Mutat 32: 564-567.
5. Green ED, Guyer MS (2011) Charting a course for genomic medicine from base pairs to bedside. Nature 470: 204-213.
6. Hindorff LA, MacArthur J, J. M, Junkins HA, Hall PN, et al. (2013) A Catalog of Published Genome-wide Association Studies.
7. Hindorff LA, Sethupathy P, Junkins HA, Ramos EM, Mehta JP, et al. (2009) Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. Proc Natl Acad Sci U S A 106: 9362-9367.
8. Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, et al. (2009) Finding the missing heritability of complex diseases. Nature 461: 747-753.
9. Maher B (2008) Personal genomes: The case of the missing heritability. Nature 456: 18-21.
10. Lockshin MD (2006) Sex differences in autoimmune disease. LUPUS 15: 753-756.
11. Whitacre CC, Reingold SC, O'Looney PA (1999) A gender gap in autoimmunity. Science 283: 1277-1278.
12. Whitacre CC (2001) Sex differences in autoimmune disease. Nat Immunol 2: 777-780.
13. Gater R, Tansella M, Korten A, Tiemens BG, Mavreas VG, 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. Arch Gen Psychiatry 55: 405-413.
14. Lai F, Kammann E, Rebeck GW, Anderson A, Chen Y, et al. (1999) APOE genotype and gender effects on Alzheimer disease in 100 adults with Down syndrome. Neurology 53: 331-336.
15. Andersen K, Launer LJ, Dewey ME, Letenneur L, Ott A, et al. (1999) Gender differences in the incidence of AD and vascular dementia: The EURODEM Studies. EURODEM Incidence Research Group. Neurology 53: 1992-1997.
16. Goldstein JM, Seidman LJ, Horton NJ, Makris N, Kennedy DN, et al. (2001) Normal sexual dimorphism of the adult human brain assessed by in vivo magnetic resonance imaging. Cereb Cortex 11: 490-497.
17. Jazin E, Cahill L (2010) Sex differences in molecular neuroscience: from fruit flies to humans. Nat Rev Neurosci 11: 9-17.
18. Choi BG, McLaughlin MA (2007) Why men's hearts break: cardiovascular effects of sex steroids. Endocrinol Metab Clin North Am 36: 365-377.

bioRxiv preprint doi: https://doi.org/10.1101/009464; this version posted September 21, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

19. Anderson KM, Odell PM, Wilson PW, Kannel WB (1991) Cardiovascular disease risk profiles. Am Heart J 121: 293-298.
20. Mendelsohn ME, Karas RH (2005) Molecular and cellular basis of cardiovascular gender differences. Science 308: 1583-1587.
21. Lerner DJ, Kannel WB (1986) Patterns of coronary heart disease morbidity and mortality in the sexes: a 26-year follow-up of the Framingham population. Am Heart J 111: 383-390.
22. Teslovich TM, Musunuru K, Smith AV, Edmondson AC, Stylianou IM, et al. (2010) Biological, clinical and population relevance of 95 loci for blood lipids. Nature 466: 707-713.
23. Matanoski G, Tao X, Almon L, Adade AA, Davies-Cole JO (2006) Demographics and tumor characteristics of colorectal cancers in the United States, 1998-2001. Cancer 107: 1112-1120.
24. Muscat JE, Richie JP, Jr., Thompson S, Wynder EL (1996) Gender differences in smoking and risk for oral cancer. Cancer Res 56: 5192-5197.
25. Naugler WE, Sakurai T, Kim S, Maeda S, Kim K, et al. (2007) Gender disparity in liver cancer due to sex differences in MyD88-dependent IL-6 production. Science 317: 121-124.
26. Zang EA, Wynder EL (1996) Differences in lung cancer risk between men and women: examination of the evidence. J Natl Cancer Inst 88: 183-192.
27. Ober C, Loisel Da, Gilad Y (2008) Sex-specific genetic architecture of human disease. Nat Rev Genet 9: 911-922.
28. Patsopoulos NA, Tatsioni A, Ioannidis JP (2007) Claims of sex differences: an empirical assessment in genetic associations. Jama 298: 880-893.
29. Fish EN (2008) The X-files in immunity: sex-based differences predispose immune responses. Nat Rev Immunol 8: 737-744.
30. Nelson JL, Ostensen M (1997) Pregnancy and rheumatoid arthritis. Rheum Dis Clin North Am 23: 195-212.
31. Confavreux C, Hutchinson M, Hours MM, Cortinovis-Tourniaire P, Moreau T (1998) Rate of pregnancy-related relapse in multiple sclerosis. Pregnancy in Multiple Sclerosis Group. New Engl J Med 339: 285-291.
32. Ross MT, Grafham DV, Coffey AJ, Scherer S, McLay K, et al. (2005) The DNA sequence of the human X chromosome. Nature 434: 325-337.
33. Carrel L, Willard HF (2005) X-inactivation profile reveals extensive variability in X-linked gene expression in females. Nature 434: 400-404.
34. Tarpey PS, Smith R, Pleasance E, Whibley A, Edkins S, et al. (2009) A systematic, large-scale resequencing screen of X-chromosome coding exons in mental retardation. Nat Genet 41: 535-543.
35. Ropers HH, Hamel BC (2005) X-linked mental retardation. Nature reviews Genetics 6: 46-57.
36. Morrow EH, Connallon T (2013) Implications of sex-specific selection for the genetic basis of disease. Evol Appl 6: 1208-1217.
37. Kemkemer C, Kohn M, Kehrer-Sawatzki H, Fundele RH, Hameister H (2009) Enrichment of brain-related genes on the mammalian X chromosome is ancient and predates the divergence of synapsid and sauropsid lineages. Chromosome Res 17: 811-820.

bioRxiv preprint doi: https://doi.org/10.1101/009464; this version posted September 21, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

38. Saifi GM, Chandra HS (1999) An apparent excess of sex- and reproduction-related genes on the human X chromosome. Proc Biol Sci 266: 203-209.
39. Nguyen DK, Disteche CM (2006) High expression of the mammalian X chromosome in brain. Brain Res 1126: 46-49.
40. Dobyns WB, Filauro A, Tomson BN, Chan AS, Ho AW, et al. (2004) Inheritance of most X-linked traits is not dominant or recessive, just X-linked. Am J Med Genet A 129A: 136-143.
41. Gottipati S, Arbiza L, Siepel A, Clark AG, Keinan A (2011) Analyses of X-linked and autosomal genetic variation in population-scale whole genome sequencing. Nat Genet 43: 741-743.
42. Lohmueller KE, Degenhardt JD, Keinan A (2010) Sex-averaged recombination and mutation rates on the X chromosome: a comment on Labuda et al. Am J Hum Genet 86: 978-980; author reply 980-971.
43. Keinan A, Reich D (2010) Can a sex-biased human demography account for the reduced effective population size of chromosome X in non-Africans? Mol Biol Evol 27: 2312-2321.
44. Keinan A, Mullikin JC, Patterson N, Reich D (2007) Measurement of the human allele frequency spectrum demonstrates greater genetic drift in East Asians than in Europeans. Nat Genet 39: 1251-1255.
45. Keinan A, Mullikin JC, Patterson N, Reich D (2009) Accelerated genetic drift on chromosome X during the human dispersal out of Africa. Nat Genet 41: 66-70.
46. Hammer MF, Woerner AE, Mendez FL, Watkins JC, Cox MP, et al. (2010) The ratio of human X chromosome to autosome diversity is positively correlated with genetic distance from genes. Nat Genet 42: 830-831.
47. Hammer MF, Mendez FL, Cox MP, Woerner AE, Wall JD (2008) Sex-biased evolutionary forces shape genomic patterns of human diversity. PLoS Genet 4: e1000202.
48. Patterson N, Price AL, Reich D (2006) Population structure and eigenanalysis. PLoS Genet 2: e190.
49. Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, et al. (2006) Principal components analysis corrects for stratification in genome-wide association studies. Nat Genet 38: 904-909.
50. Novembre J, Johnson T, Bryc K, Kutalik Z, Boyko AR, et al. (2008) Genes mirror geography within Europe. Nature 456: 98-101.
51. Zheng G, Joo J, Zhang C, Geller NL (2007) Testing association for markers on the X chromosome. Genet Epidemiol 31: 834-843.
52. Clayton DG (2009) Sex chromosomes and genetic association studies. Genome Med 1: 110.
53. Clayton D (2008) Testing for association on the X chromosome. Biostatistics 9: 593-600.
54. Thornton T, Zhang Q, Cai X, Ober C, McPeek MS (2012) XM: Association Testing on the X-Chromosome in Case-Control Samples With Related Individuals. Genet Epidemiol.
55. Purcell S, Neale B, Toddbrown K, Thomas L, Ferreira M, et al. (2007) PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. Am J Hum Genet 81: 559-575.
56. Tukiainen T, Pirinen M, Sarin AP, Ladenvall C, Kettunen J, et al. (2014) Chromosome x-wide association study identifies Loci for fasting insulin and height and evidence for incomplete dosage compensation. PLoS Genet 10: e1004127.

bioRxiv preprint doi: https://doi.org/10.1101/009464; this version posted September 21, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

57. Loley C, Ziegler A, Konig IR (2011) Association tests for X-chromosomal markers--a comparison of different test statistics. Hum Hered 71: 23-36.
58. Gleicher N, Barad DH (2007) Gender as risk factor for autoimmune diseases. J Autoimmun 28: 1-6.
59. Beeson PB (1994) Age and sex associations of 40 autoimmune diseases. Am J Med 96: 457-462.
60. Sawalha AH, Webb R, Han S, Kelly JA, Kaufman KM, et al. (2008) Common variants within MECP2 confer risk of systemic lupus erythematosus. PLoS One 3: e1727.
61. Shen N, Fu Q, Deng Y, Qian X, Zhao J, et al. (2010) Sex-specific association of X-linked Toll-like receptor 7 (TLR7) with male systemic lupus erythematosus. Proc Natl Acad Sci U S A 107: 15838-15843.
62. Selmi C, Brunetta E, Raimondo MG, Meroni PL (2012) The X chromosome and the sex ratio of autoimmunity. Autoimmun Rev 11: A531-537.
63. Tiniakou E, Costenbader KH, Kriegel MA (2013) Sex-specific environmental influences on the development of autoimmune diseases. Clin Immunol 149: 182-191.
64. Quintero OL, Amador-Patarroyo MJ, Montoya-Ortiz G, Rojas-Villarraga A, Anaya JM (2012) Autoimmune disease and gender: plausible mechanisms for the female predominance of autoimmunity. J Autoimmun 38: J109-119.
65. Libert C, Dejager L, Pinheiro I (2010) The X chromosome in immune functions: when a chromosome makes the difference. Nat Rev Immunol 10: 594-604.
66. Bianchi I, Lleo A, Gershwin ME, Invernizzi P (2011) The X chromosome and immune associated genes. J Autoimmun.
67. Jostins L, Ripke S, Weersma RK, Duerr RH, McGovern DP, et al. (2012) Host-microbe interactions have shaped the genetic architecture of inflammatory bowel disease. Nature 491: 119-124.
68. Tysk C, Lindberg E, Jarnerot G, Floderus-Myrhed B (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: 990-996.
69. Sofaer J (1993) Crohn's disease: the genetic contribution. Gut 34: 869-871.
70. Itariu BK, Stulnig TM (2014) Autoimmune Aspects of Type 2 Diabetes Mellitus - A Mini-Review. Gerontology.
71. Pagani MR, Gonzalez LE, Uchitel OD (2011) Autoimmunity in amyotrophic lateral sclerosis: past and present. Neurol Res Int 2011: 497080.
72. 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.
73. Liu JZ, McRae AF, Nyholt DR, Medland SE, Wray NR, et al. (2010) A versatile gene-based test for genome-wide association studies. Am J Hum Genet 87: 139-145.
74. Neale BM, Sham PC (2004) The future of association studies: gene-based analysis and replication. Am J Hum Genet 75: 353-362.
75. Beyene J, Tritchler D, Asimit JL, Hamid JS (2009) Gene- or region-based analysis of genome-wide association studies. Genet Epidemiol 33 Suppl 1: S105-110.
76. Li MX, Gui HS, Kwan JS, Sham PC (2011) GATES: a rapid and powerful gene-based association test using extended Simes procedure. Am J Hum Genet 88: 283-293.
77. Jiang B, Zhang X, Zuo Y, Kang G (2011) A powerful truncated tail strength method for testing multiple null hypotheses in one dataset. J Theor Biol 277: 67-73.

bioRxiv preprint doi: https://doi.org/10.1101/009464; this version posted September 21, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

78. Zaykin DV, Zhivotovsky LA, Westfall PH, Weir BS (2002) Truncated product method for combining P-values. Genet Epidemiol 22: 170-185.
79. Ma L, Clark AG, Keinan A (2013) Gene-Based Testing of Interactions in Association Studies of Quantitative Traits. PLoS Genet 9.
80. Huang HL, Chanda P, Alonso A, Bader JS, Arking DE (2011) Gene-Based Tests of Association. PLoS Genet 7.
81. Sirota M, Schaub Ma, Batzoglou S, Robinson WH, Butte AJ (2009) Autoimmune disease classification by inverse association with SNP alleles. PLoS Genet 5: e1000792.
82. Cotsapas C, Voight BF, Rossin E, Lage K, Neale BM, et al. (2011) Pervasive sharing of genetic effects in autoimmune disease. PLoS Genet 7: e1002254.
83. Sivakumaran S, Agakov F, Theodoratou E, Prendergast JG, Zgaga L, et al. (2011) Abundant pleiotropy in human complex diseases and traits. Am J Hum Genet 89: 607-618.
84. Chang D, Keinan A (2014) Principal component analysis characterizes shared pathogenetics from genome-wide association studies. in press.
85. Birlea SA, Jin Y, Bennett DC, Herbstman DM, Wallace MR, et al. (2011) Comprehensive association analysis of candidate genes for generalized vitiligo supports XBP1, FOXP3, and TSLP. J Invest Dermatol 131: 371-381.
86. Tang QZ, Bluestone JA (2008) The Foxp3(+) regulatory T cell: a jack of all trades, master of regulation. Nat Immunol 9: 239-244.
87. Fontenot JD, Gavin MA, Rudensky AY (2003) Foxp3 programs the development and function of CD4(+)CD25(+) regulatory T cells. Nat Immunol 4: 330-336.
88. Bennett CL, Christie J, Ramsdell F, Brunkow ME, Ferguson PJ, et al. (2001) The immune dysregulation, polyendocrinopathy, enteropathy, X-linked syndrome (IPEX) is caused by mutations of FOXP3. Nat Genet 27: 20-21.
89. Baek HY, Lim JW, Kim H (2007) Interaction between the Helicobacter pylori CagA and alpha-Pix in gastric epithelial AGS cells. Ann N Y Acad Sci 1096: 18-23.
90. Luther J, Dave M, Higgins PD, Kao JY (2010) Association between Helicobacter pylori infection and inflammatory bowel disease: a meta-analysis and systematic review of the literature. Inflamm Bowel Dis 16: 1077-1084.
91. Jin X, Chen YP, Chen SH, Xiang Z (2013) Association between Helicobacter Pylori infection and ulcerative colitis--a case control study from China. Int J Med Sci 10: 1479-1484.
92. Matson DR, Demirel PB, Stukenberg PT, Burke DJ (2012) A conserved role for COMA/CENP-H/I/N kinetochore proteins in the spindle checkpoint. Genes Dev 26: 542-547.
93. Hamdouch K, Rodriguez C, Perez-Venegas J, Rodriguez I, Astola A, et al. (2011) Anti-CENPI autoantibodies in scleroderma patients with features of autoimmune liver diseases. Clin Chim Acta 412: 2267-2271.
94. Baranzini SE, Wang J, Gibson RA, Galwey N, Naegelin Y, et al. (2009) Genome-wide association analysis of susceptibility and clinical phenotype in multiple sclerosis. Hum Mol Genet 18: 767-778.
95. Ahmeti KB, Ajroud-Driss S, Al-Chalabi A, Andersen PM, Armstrong J, et al. (2013) Age of onset of amyotrophic lateral sclerosis is modulated by a locus on 1p34.1. Neurobiol Aging 34: 357 e357-319.

bioRxiv preprint doi: https://doi.org/10.1101/009464; this version posted September 21, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

96. Ju T, Cummings RD (2005) Protein glycosylation: chaperone mutation in Tn syndrome. Nature 437: 1252.
97. Thurnher M, Clausen H, Fierz W, Lanzavecchia A, Berger EG (1992) T cell clones with normal or defective O-galactosylation from a patient with permanent mixed-field polyagglutinability. Eur J Immunol 22: 1835-1842.
98. Wilson Sayres MA, Makova KD (2013) Gene survival and death on the human Y chromosome. Mol Biol Evol 30: 781-787.
99. Sharp A, Robinson D, Jacobs P (2000) Age- and tissue-specific variation of X chromosome inactivation ratios in normal women. Hum Genet 107: 343-349.
100. Cotton AM, Lam L, Affleck JG, Wilson IM, Penaherrera MS, et al. (2011) Chromosome-wide DNA methylation analysis predicts human tissue-specific X inactivation. Hum Genet 130: 187-201.
101. Calvo JR, Gonzalez-Yanes C, Maldonado MD (2013) The role of melatonin in the cells of the innate immunity: a review. J Pineal Res 55: 103-120.
102. Pohanka M (2013) Impact of melatonin on immunity: a review. Cent Eur J Med 8: 369-376.
103. Dibner C, Schibler U, Albrecht U (2010) The Mammalian Circadian Timing System: Organization and Coordination of Central and Peripheral Clocks. Annu Rev Physiol 72: 517-549.
104. Jacob S, Poeggeler B, Weishaupt JH, Siren AL, Hardeland R, 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. J Pineal Res 33: 186-187.
105. Terry PD, Villinger F, Bubenik GA, Sitaraman SV (2009) Melatonin and Ulcerative Colitis: Evidence, Biological Mechanisms, and Future Research. Inflamm Bowel Dis 15: 134-140.
106. Slominski A, Paus R, Bomirski A (1989) Hypothesis - Possible Role for the Melatonin Receptor in Vitiligo - Discussion Paper. J R Soc Med 82: 539-541.
107. Sospedra M, Martin R (2005) Immunology of multiple sclerosis. Annu Rev Immunol 23: 683-747.
108. Weishaupt JH, Bartels C, Polking E, Dietrich J, Rohde G, et al. (2006) Reduced oxidative damage in ALS by high-dose enteral melatonin treatment. J Pineal Res 41: 313-323.
109. Mahan AL, Ressler KJ (2012) Fear conditioning, synaptic plasticity and the amygdala: implications for posttraumatic stress disorder. Trends Neurosci 35: 24-35.
110. Roozendaal B, McEwen BS, Chattarji S (2009) Stress, memory and the amygdala. Nature Reviews Neuroscience 10: 423-433.
111. Heller MM, Lee ES, Koo JY (2011) Stress as an influencing factor in psoriasis. Skin Therapy Lett 16: 1-4.
112. Schuster N, Krieglstein K (2002) Mechanisms of TGF-beta-mediated apoptosis. Cell Tissue Res 307: 1-14.
113. Lukashev ME, Werb Z (1998) ECM signalling: orchestrating cell behaviour and misbehaviour. Trends Cell Biol 8: 437-441.
114. Logan CY, Nusse R (2004) The Wnt signaling pathway in development and disease. Annu Rev Cell Dev Biol 20: 781-810.
115. Staal FJ, Luis TC, Tiemessen MM (2008) WNT signalling in the immune system: WNT is spreading its wings. Nat Rev Immunol 8: 581-593.

bioRxiv preprint doi: https://doi.org/10.1101/009464; this version posted September 21, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

116. Eguchi K (2001) Apoptosis in autoimmune diseases. Intern Med 40: 275-284.
117. Kawakami A, Eguchi K (2002) Involvement of apoptotic cell death in autoimmune diseases. Med Electron Microsc 35: 1-8.
118. Mason KD, Lin A, Robb L, Josefsson EC, Henley KJ, et al. (2013) Proapoptotic Bak and Bax guard against fatal systemic and organ-specific autoimmune disease. Proc Natl Acad Sci U S A 110: 2599-2604.
119. Moretti S, Fabbri P, Baroni G, Berti S, Bani D, et al. (2009) Keratinocyte dysfunction in vitiligo epidermis: cytokine microenvironment and correlation to keratinocyte apoptosis. Histol Histopathol 24: 849-857.
120. Weatherhead SC, Farr PM, Jamieson D, Hallinan JS, Lloyd JJ, et al. (2011) Keratinocyte apoptosis in epidermal remodeling and clearance of psoriasis induced by UV radiation. J Invest Dermatol 131: 1916-1926.
121. Li N, Ma T, Han J, Zhou J, Wang J, 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. Eur Rev Med Pharmacol Sci 18: 78-85.
122. Konig IR, Loley C, Erdmann J, Ziegler A (2014) How to include chromosome x in your genome-wide association study. Genet Epidemiol 38: 97-103.
123. Conde L, Foo JN, Riby J, Liu J, Darabi H, et al. (2013) X chromosome-wide association study of follicular lymphoma. Br J Haematol 162: 858-862.
124. Chang D, Gao F, Keinan A XWAS: a toolset for genetic data analysis and association studies of the X chromosome. Under Review.
125. Laaksovirta H, Peuralinna T, Schymick JC, Scholz SW, Lai SL, et al. (2010) Chromosome 9p21 in amyotrophic lateral sclerosis in Finland: a genome-wide association study. Lancet Neurol 9: 978-985.
126. Cronin S, Berger S, Ding J, Schymick JC, Washecka N, et al. (2008) A genome-wide association study of sporadic ALS in a homogenous Irish population. Hum Mol Genet 17: 768-774.
127. Ahn R, Ding YC, Murray J, Fasano A, Green PH, et al. (2012) Association analysis of the extended MHC region in celiac disease implicates multiple independent susceptibility loci. PLoS One 7: e36926.
128. Jin Y, Birlea SA, Fain PR, Gowan K, Riccardi SL, et al. (2010) Variant of TYR and autoimmunity susceptibility loci in generalized vitiligo. New Engl J Med 362: 1686-1697.
129. Duerr RH, Taylor KD, Brant SR, Rioux JD, Silverberg MS, et al. (2006) A genome-wide association study identifies IL23R as an inflammatory bowel disease gene. Science 314: 1461-1463.
130. Nair RP, Duffin KC, Helms C, Ding J, Stuart PE, et al. (2009) Genome-wide scan reveals association of psoriasis with IL-23 and NF-kappaB pathways. Nat Genet 41: 199-204.
131. Qi L, Cornelis MC, Kraft P, Stanya KJ, Linda Kao WH, et al. (2010) Genetic variants at 2q24 are associated with susceptibility to type 2 diabetes. Hum Mol Genet 19: 2706-2715.
132. Evans DM, Spencer CC, Pointon JJ, Su Z, Harvey D, et al. (2011) Interaction between ERAP1 and HLA-B27 in ankylosing spondylitis implicates peptide handling in the mechanism for HLA-B27 in disease susceptibility. Nat Genet 43: 761-767.

bioRxiv preprint doi: https://doi.org/10.1101/009464; this version posted September 21, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

133. Barrett JC, Lee JC, Lees CW, Prescott NJ, Anderson CA, et al. (2009) Genome-wide association study of ulcerative colitis identifies three new susceptibility loci, including the HNF4A region. Nat Genet 41: 1330-1334.
134. Sawcer S, Hellenthal G, Pirinen M, Spencer CC, Patsopoulos NA, et al. (2011) Genetic risk and a primary role for cell-mediated immune mechanisms in multiple sclerosis. Nature 476: 214-219.
135. Jin Y, Birlea SA, Fain PR, Ferrara TM, Ben S, et al. (2012) Genome-wide association analyses identify 13 new susceptibility loci for generalized vitiligo. Nat Genet 44: 676-680.
136. Petersen GM, Amundadottir L, Fuchs CS, Kraft P, Stolzenberg-Solomon RZ, et al. (2010) A genome-wide association study identifies pancreatic cancer susceptibility loci on chromosomes 13q22.1, 1q32.1 and 5p15.33. Nat Genet 42: 224-228.
137. Amundadottir L, Kraft P, Stolzenberg-Solomon RZ, Fuchs CS, Petersen GM, et al. (2009) Genome-wide association study identifies variants in the ABO locus associated with susceptibility to pancreatic cancer. Nat Genet 41: 986-990.
138. Lee JH, Cheng R, Graff-Radford N, Foroud T, Mayeux R (2008) Analyses of the National Institute on Aging Late-Onset Alzheimer's Disease Family Study: implication of additional loci. Arch Neurol 65: 1518-1526.
139. Estrada K, Styrkarsdottir U, Evangelou E, Hsu YH, Duncan EL, et al. (2012) Genome-wide meta-analysis identifies 56 bone mineral density loci and reveals 14 loci associated with risk of fracture. Nat Genet 44: 491-501.
140. Bierut LJ, Saccone NL, Rice JP, Goate A, Foroud T, et al. (2002) Defining alcohol-related phenotypes in humans. The Collaborative Study on the Genetics of Alcoholism. Alcohol Res Health 26: 208-213.
141. Bierut LJ, Strickland JR, Thompson JR, Afful SE, Cottler LB (2008) Drug use and dependence in cocaine dependent subjects, community-based individuals, and their siblings. Drug Alcohol Depend 95: 14-22.
142. Bierut LJ (2007) Genetic variation that contributes to nicotine dependence. Pharmacogenomics 8: 881-883.
143. Laurie CC, Doheny KF, Mirel DB, Pugh EW, Bierut LJ, et al. (2010) Quality control and quality assurance in genotypic data for genome-wide association studies. Genet Epidemiol 34: 591-602.
144. Ling H, Hetrick K, Bailey-Wilson JE, Pugh EW (2009) Application of sex-specific single-nucleotide polymorphism filters in genome-wide association data. BMC Proc 3 Suppl 7: S57.
145. Ziegler A (2009) Genome-wide association studies: quality control and population-based measures. Genet Epidemiol 33 Suppl 1: S45-50.
146. Heyer E, Chaix R, Pavard S, Austerlitz F (2012) Sex-specific demographic behaviours that shape human genomic variation. Mol Ecol 21: 597-612.
147. Wilder JA, Kingan SB, Mobasher Z, Pilkington MM, Hammer MF (2004) Global patterns of human mitochondrial DNA and Y-chromosome structure are not influenced by higher migration rates of females versus males. Nat Genet 36: 1122-1125.
148. Oota H, Settheetham-Ishida W, Tiwawech D, Ishida T, Stoneking M (2001) Human mtDNA and Y-chromosome variation is correlated with matrilocal versus patrilocal residence. Nat Genet 29: 20-21.

bioRxiv preprint doi: https://doi.org/10.1101/009464; this version posted September 21, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

149. Howie BN, Donnelly P, Marchini J (2009) A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. PLoS Genet 5: e1000529.
150. Abecasis GR, Auton A, Brooks LD, DePristo MA, Durbin RM, et al. (2012) An integrated map of genetic variation from 1,092 human genomes. Nature 491: 56-65.
151. Fisher RA (1925) Statistical Methods for Research Workers. Edinburgh: Oliver and Boyd.
152. Stouffer SA, Suchman EA, Devinney LC, Star SA, Williams RMJ (1949) Adjustment During Army Life. Princeton, NJ: Princeton University Press.
153. Willer CJ, Li Y, Abecasis GR (2010) METAL: fast and efficient meta-analysis of genomewide association scans. Bioinformatics 26: 2190-2191.
154. Randall JC, Winkler TW, Kutalik Z, Berndt SI, Jackson AU, et al. (2013) Sex-stratified genome-wide association studies including 270,000 individuals show sexual dimorphism in genetic loci for anthropometric traits. PLoS Genet 9: e1003500.
155. Hofman A, Breteler MM, van Duijn CM, Janssen HL, Krestin GP, et al. (2009) The Rotterdam Study: 2010 objectives and design update. Eur J Epidemiol 24: 553-572.
156. Barrett T, Wilhite SE, Ledoux P, Evangelista C, Kim IF, et al. (2013) NCBI GEO: archive for functional genomics data sets--update. Nucleic Acids Res 41: D991-995.
157. Su AI, Wiltshire T, Batalov S, Lapp H, Ching KA, et al. (2004) A gene atlas of the mouse and human protein-encoding transcriptomes. Proc Natl Acad Sci U S A 101: 6062-6067.
158. Wu C, Macleod I, Su AI (2013) BioGPS and MyGene.info: organizing online, gene-centric information. Nucleic Acids Res 41: D561-565.
159. Warde-Farley D, Donaldson SL, Comes O, Zuberi K, Badrawi R, et al. (2010) The GeneMANIA prediction server: biological network integration for gene prioritization and predicting gene function. Nucleic Acids Res 38: W214-220.
160. Wang J, Duncan D, Shi Z, Zhang B (2013) WEB-based GEne SeT AnaLysis Toolkit (WebGestalt): update 2013. Nucleic Acids Res.
161. Zhang B, Kirov S, Snoddy J (2005) WebGestalt: an integrated system for exploring gene sets in various biological contexts. Nucleic Acids Res 33: W741-748.
162. Kanehisa M, Goto S (2000) KEGG: kyoto encyclopedia of genes and genomes. Nucleic Acids Res 28: 27-30.
163. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, et al. (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet 25: 25-29.
164. Thomas PD, Campbell MJ, Kejariwal A, Mi H, Karlak B, et al. (2003) PANTHER: a library of protein families and subfamilies indexed by function. Genome Res 13: 2129-2141.

bioRxiv preprint doi: https://doi.org/10.1101/009464; this version posted September 21, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

FIGURE LEGENDS

Figure 1. X-linked genes associated with autoimmune disease risk. All genes that showed evidence of association in a gene-based test, and replicated in any other dataset we analyzed (see main text) are presented for the a) $FM_{S.comb}$ b) $FM_{F.comb}$ c) FM_{02} and d) sex-differentiated effect size tests (Materials and Methods). *X-axis* denotes the different datasets, with their names following the notation from Table 1, which are the same across all panels. *Y-axis* displays the different gene names. 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 and "**" indicates datasets in which replication is significant after correcting for the number of genes tested for replication. 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 Tables 2-3.

Figure 2. X-linked autoimmune 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 z-scores across all 504 X-linked genes that were tested as part of the entire XWAS for which expression data was available. Several associated genes exhibit significantly higher expression in immune-related tissues (see main text and Figure 3).

Figure 3. Three X-linked disease risk genes show high expression in immune-related tissues and cells. *ARHGEF6* (a), *IL13RA1* (b), and *ITM2A* (c) 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 (Figure 2). *Y-axis* follows the respective tissues from Figure 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 4. Interactome of X-linked disease risk genes. All 22 X-linked protein-coding genes that showed evidence of association, and suggestively replicated in any dataset (Figure 1), denoted by black diamonds, together with genes that interact with them (Materials and Methods). *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 [159]. All but four of these 22 genes share interacting partners according to these known and predicted interactions. Results of a pathway analysis based on this interactome are presented in Table 5.

TABLES

Dataset	Disease	# SNPs	# Genes (#SNPs in genes)	# Cases (males, females)	# Controls (males, females)
ALS Finland [125]	Amyotrophic Lateral Sclerosis (ALS)	207,947	970 (72,219)	400 (198, 202)	490 (103, 387)
ALS Irish [126]	Amyotrophic Lateral Sclerosis (ALS)	219,300	967 (77,043)	221 (119, 102)	210 (112, 98)
Psoriasis CASP [130]	Psoriasis	184,246	953 (62,106)	1,209 (588, 621)	1,271 (585, 686)
Celiac Disease CIDR [127]	Celiac Disease	187,284	962 (64,836)	1,576 (447, 1129)	504 (225, 279)
CD NIDDK [129]	Crohn's Disease (CD)	176,072	837 (58,874)	791 (378, 413)	922 (457, 465)
CD WT1* [72]	Crohn's Disease (CD)	150,275	930 (49,017)	1,592 (607, 985)	1,701 (923, 778)
UC WT2* [133]	Ulcerative Colitis (UC)	196,781	963 (67,422)	2,341 (1148, 1193)	1,699 (843, 856)
MS case control [94]	Multiple Sclerosis (MS)	183,954	842 (61,119)	943 (312, 631)	851 (290, 561)
MS WT2* [134]	Multiple Sclerosis (MS)	169,707	962 (58,463)	2,666 (698, 1968)	1389 (700, 689)
Vitiligo GWAS1 [128]	Vitiligo	157,676	958 (54,384)	1,391 (411, 980)	4,521 (1985, 2536)
Vitiligo GWAS2 [135]	Vitiligo	187,688	962 (64,974)	415 (144, 271)	2,552 (973, 1579)
T2D GENEVA [131]	Type-2 Diabetes (T2D)	220,752	971 (75,941)	2,515 (1050, 1465)	2,850 (1187, 1663)
T2D WT1* [72]	Type-2 Diabetes (T2D)	152,996	927 (49,956)	1,811 (1051, 760)	1,668 (710, 958)
T1D WT1* [72]	Type-1 Diabetes (T1D)	152,304	926 (49,718)	1,867 (954, 913)	1,714 (941, 773)
RA WT1* [72]	Rheumatoid Arthritis (RA)	146,907	925 (47,880)	1,772 (443, 1329)	1,709 (920, 789)

bioRxiv preprint doi: https://doi.org/10.1101/009464; this version posted September 21, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

AS WT2* [132]	Ankyolosing Spondylitis (AS)	200,042	966 (69,441)	1,472 (976, 496)	1,260 (665, 595)

Table 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 and the number of SNPs mapped to the genes or within 15kb of them [*# Genes (# SNPs in genes*)]. The number of individuals *(# Cases* and *# Controls*) represents the number of samples following QC. The number of males and females in each category is denoted in parenthesis. All datasets consist of individuals of European ancestry.
*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*.

bioRxiv preprint doi: https://doi.org/10.1101/009464; this version posted September 21, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

bioRxiv preprint doi: https://doi.org/10.1101/009464; this version posted September 21, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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 2. 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.

bioRxiv preprint doi: https://doi.org/10.1101/009464; this version posted September 21, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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	$5.10x10^{-2}$, $4.90x10^{-2}$	$4.98x10^{-4}$, $6.66x10^{-3}$
Celiac CIDR	CENPI	$2.90x10^{-3}$, $5.23x10^{-4}$	ALS Finland	$1.12x10^{-2}$, $1.00x10^{-3}$	$3.68x10^{-4}$, $8.09x10^{-6}$
			ALS Irish	$2.68x10^{-2}$, $1.64x10^{-2}$	$8.13x10^{-4}$, $1.09x10^{-4}$
			Vitiligo GWAS1	$1.55x10^{-4}$, $2.60x10^{-3}$	$7.02x10^{-6}$, $1.97x10^{-5}$
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	$6.30x10^{-3}$, $6.40x10^{-3}$	$1.35x10^{-4}$, $4.60x10^{-5}$

bioRxiv preprint doi: https://doi.org/10.1101/009464; this version posted September 21, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

			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	$3.03x10^{-3}$, $8.55x10^{-4}$	T2D WT1	$4.42x10^{-2}$, $4.68x10^{-2}$	$1.32x10^{-3}$, $4.45x10^{-4}$
T2D GENEVA	SNORA35	$2.12x10^{-3}$, $4.54x10^{-4}$	AS WT2	$2.40x10^{-3}$, $6.70x10^{-3}$	$6.71x10^{-5}$, $4.17x10^{-5}$
	IL13RA1	$6.35x10^{-3}$, $8.59x10^{-4}$	AS WT2	$6.20x10^{-3}$, $7.20x10^{-3}$	$4.39x10^{-4}$, $8.04x10^{-5}$
T2D WT1	MAGEC1	$2.63x10^{-2}$, $6.80x10^{-5}$	MS case control	$1.00x10^{-2}$, $1.54x10^{-2}$	$2.43x10^{-3}$, $1.55x10^{-5}$
Sex difference					
ALS Finland	MAGEE2	$6.5x10^{-4}$, $1.94x10^{-3}$	Vitiligo GWAS1	$3.08x10^{-2}$, $1.64x10^{-2}$	$2.37x10^{-4}$, $3.61x10^{-4}$
	NDP	$1.41x10^{-3}$, $9.34x10^{-4}$	CD WT1	$8.60x10^{-3}$, $1.33x10^{-2}$	$1.49x10^{-4}$, $1.53x10^{-4}$
CASP	NLGN4X	$2.34x10^{-4}$, $1.65x10^{-2}$	Vitiligo GWAS1	$4.52x10^{-2}$, $4.33x10^{-2}$	$1.32x10^{-4}$, $5.89x10^{-3}$
Celiac CIDR	CENPI	$4.4x10^{-3}$, $2.08x10^{-4}$	ALS Finland	$2.03x10^{-2}$, $1.78x10^{-2}$	$9.22x10^{-4}$, $5.00x10^{-5}$
			ALS Irish	$9.80x10^{-3}$, $4.40x10^{-3}$	$4.88x10^{-4}$, $1.36x10^{-5}$
Vitiligo GWAS1	BEND2	$3.99x10^{-3}$, $1.28x10^{-4}$	MS case control	$4.60x10^{-2}$, $5.20x10^{-2}$	$1.76x10^{-3}$, $8.60x10^{-5}$
Vitiligo GWAS2	MCF2	$7.00x10^{-4}$, $1.93x10^{-3}$	MS WT2	$2.38x10^{-2}$, $2.12x10^{-2}$	$2.00x10^{-4}$, $4.54x10^{-4}$
T2D GENEVA	EFHC2	$6.09x10^{-4}$, $1.12x10^{-3}$	RA WT1	$1.58x10^{-2}$, $1.40x10^{-3}$	$1.21x10^{-4}$, $2.42x10^{-5}$
RA WT1	MIR320D2	$8.69x10^{-3}$, $5.68x10^{-4}$	ALS Irish	$2.39x10^{-2}$, $2.64x10^{-2}$	$1.97x10^{-3}$, $1.82x10^{-4}$

Table 3. Gene-based associations replicating in other diseases. This table lists genes with nominal $P < 1x10^{-3}$ that replicated in a disease of a different phenotype (Methods). Combined p-values were calculated using Fisher's method.

bioRxiv preprint doi: https://doi.org/10.1101/009464; this version posted September 21, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

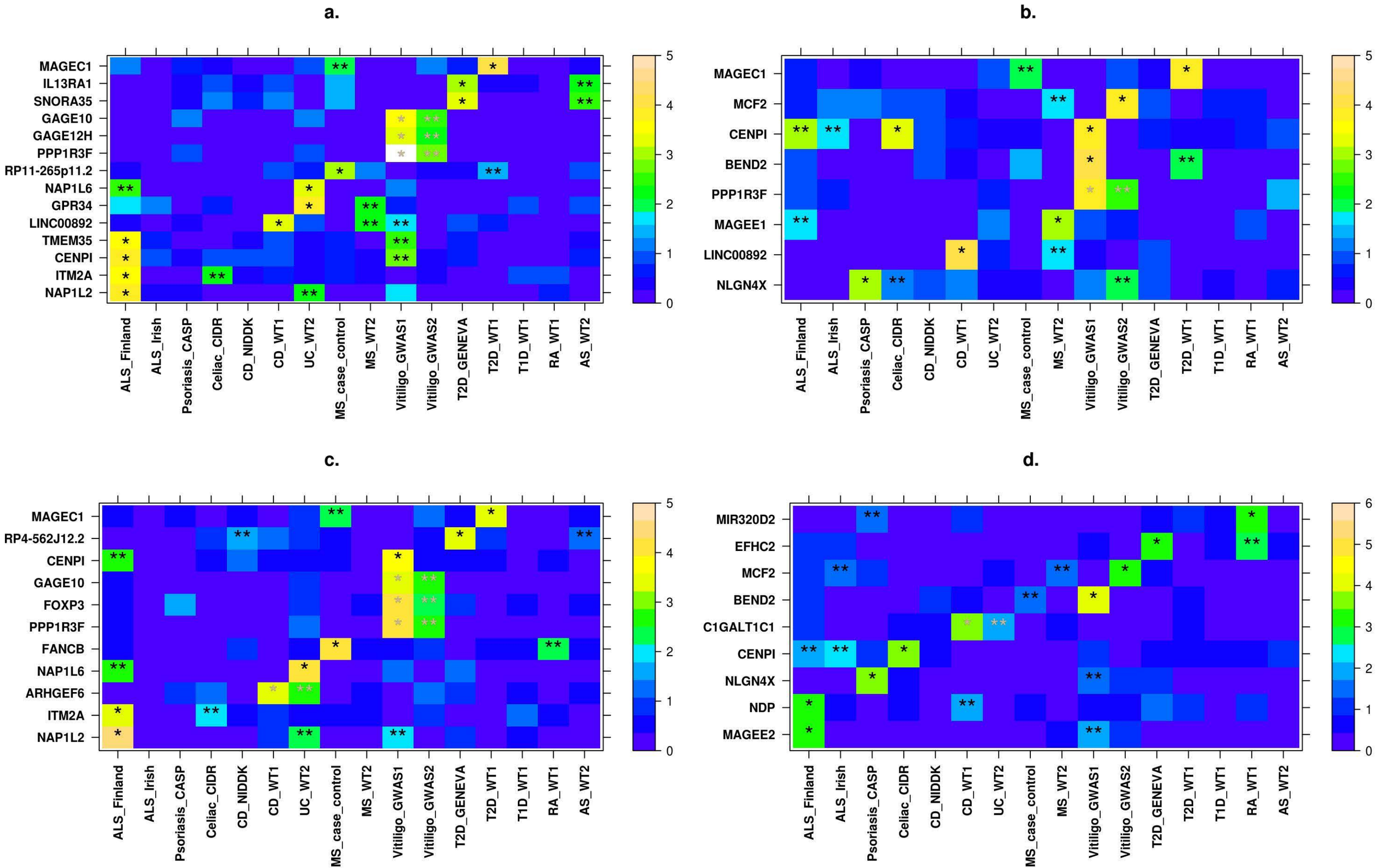
Dataset	Statistic	P-value
XY homologs gene set		
Psoriasis CASP	$FM_{F.comb}$	**0.0088**
Celiac disease CIDR	$FM_{F.comb}$	0.0467
Vitiligo GWAS1	$FM_{F.comb}$	**0.0063**
Vitiligo GWAS1	FM_{02}	0.0329
Vitiligo GWAS2	$FM_{F.comb}$	0.0346
CD NIDDK	FM_{02}	0.017
CD WT1	FM_{02}	0.0234
T1D WT1	$FM_{S.comb}$	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}**

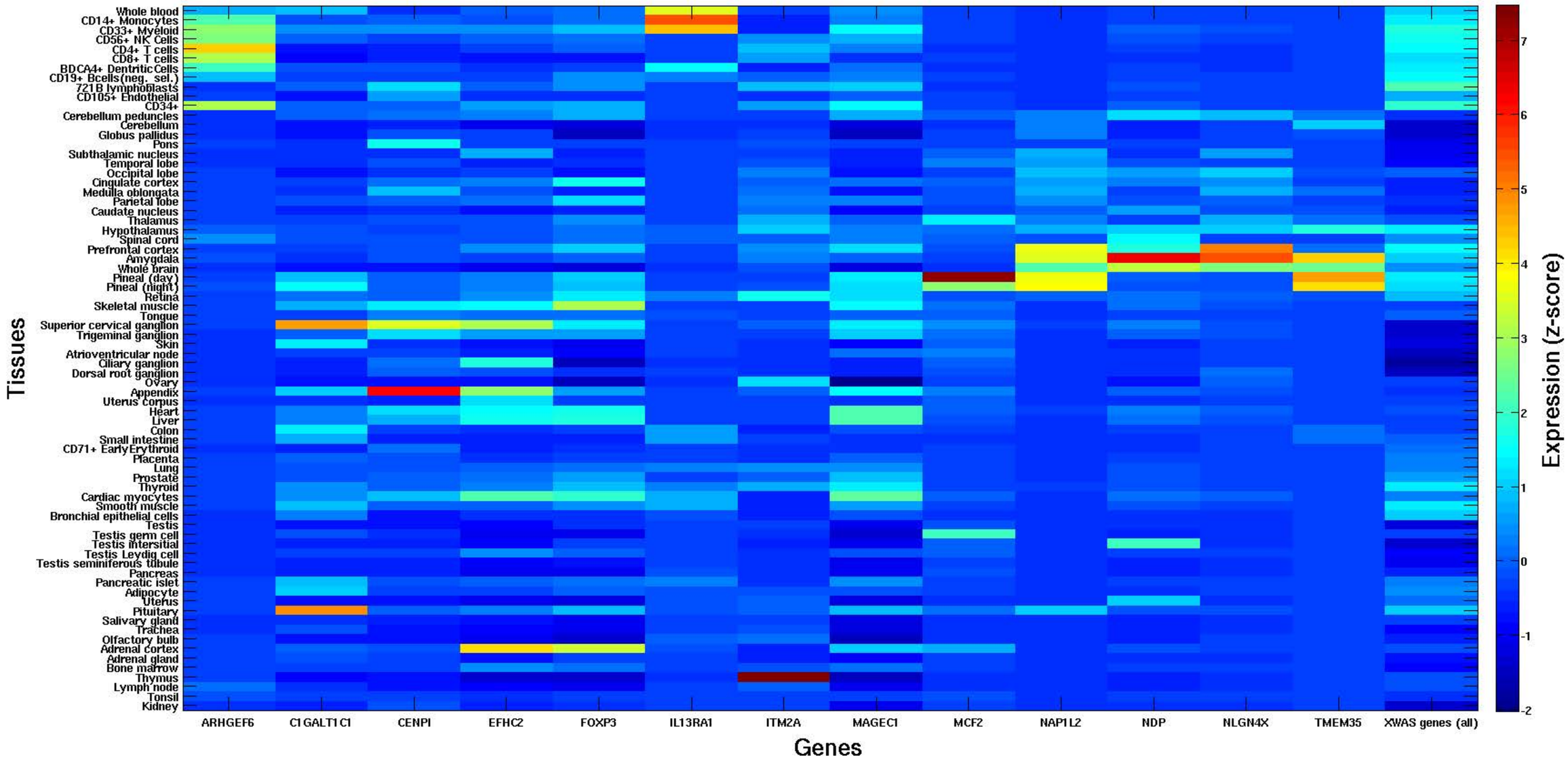
Table 4. Gene set associations. Three curated gene sets were tested for association to diseases. Displayed are only datasets with p-value < 0.05 for association with the indicated gene set, with bold p-values indicating significant association after multiple testing correction. The minimum of the truncated tail strength method and the truncated product method are displayed. Results for all datasets and tests are presented in Table S8.

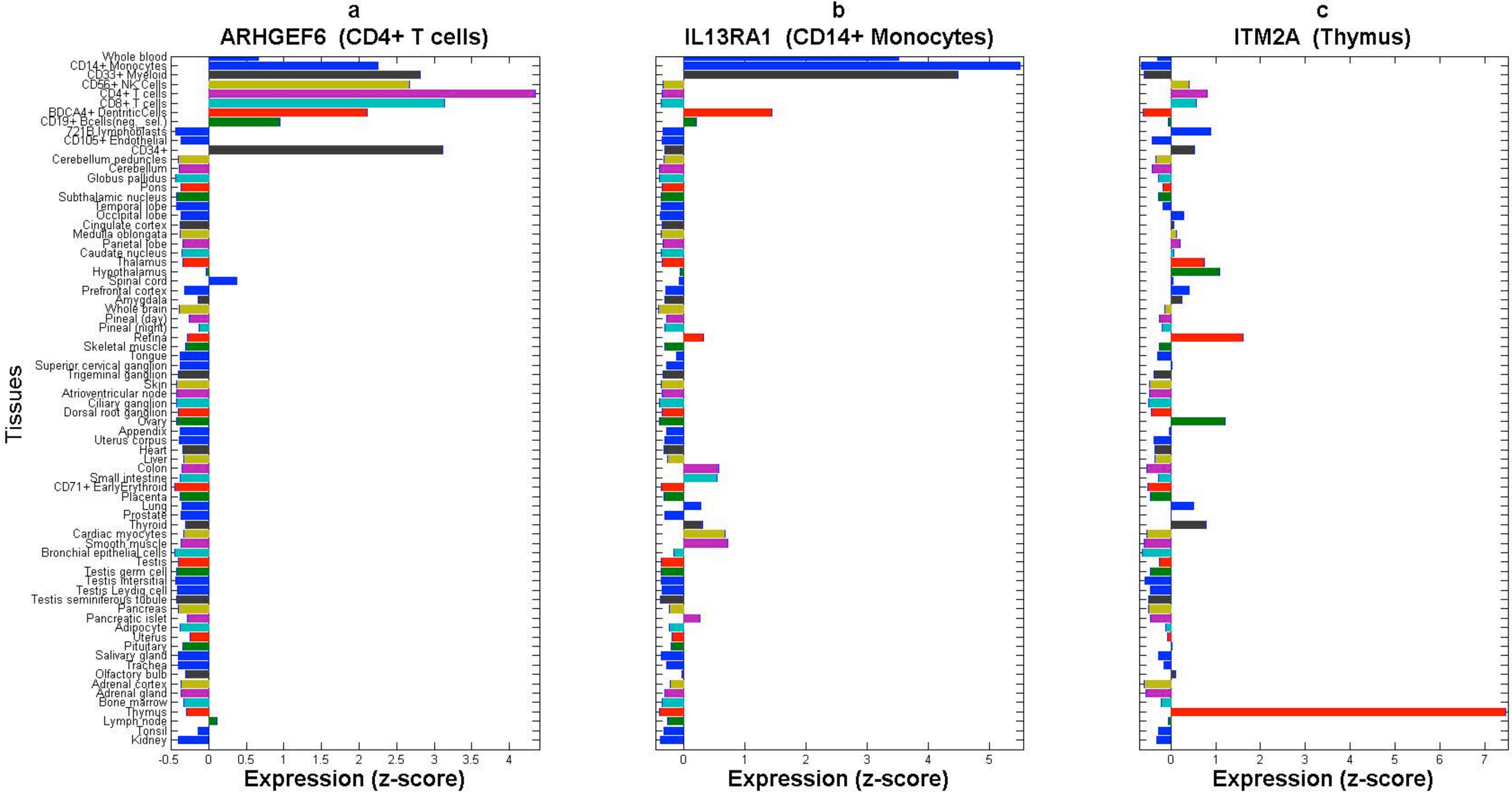
Pathway	Genes	P-value
Regulation of actin cytoskeleton	*PAK1, RHOA, PAK3, CDC42, ARHGEF6, SOS1, ARHGEF7, PAK2, RDX, GIT1, GNA13, TIAM1, ROCK2, FGD1*	5.55×10^{-14}
T-cell receptor signaling pathway	*PAK1, RHOA, PAK3, CDC42, SOS1, PAK2, IL4, NFATC2, NFATC1, ICOS, NFAT5*	2.75×10^{-13}
Axon guidance	*PAK1, 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}

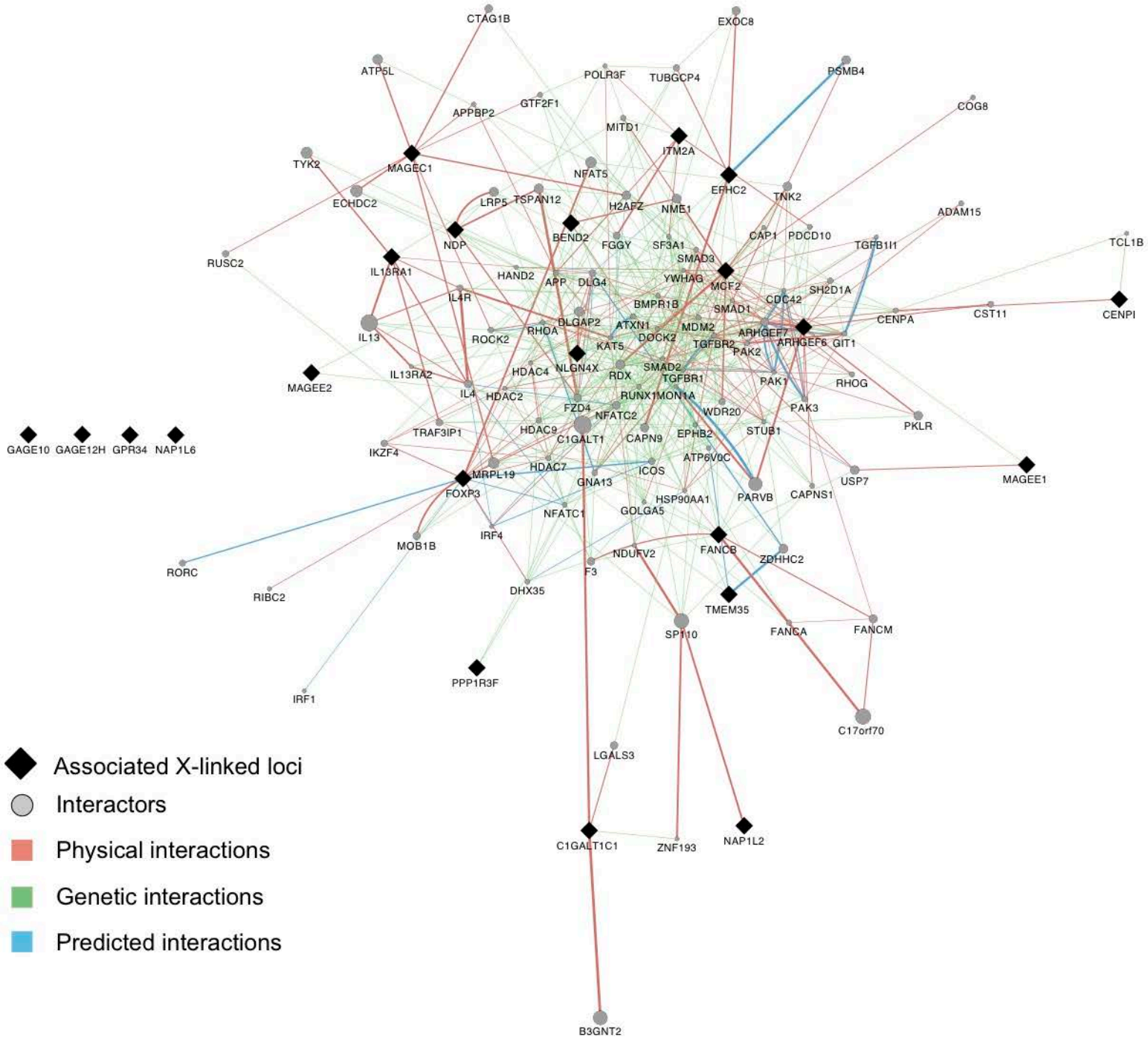
Table 5. Gene-enrichment analysis of the interactome. Genes associated to AID and DPACs, and their interacting partners (Figure 4) were enriched for several immune related pathways. We display the ten most significantly enriched pathways. Genes within each pathway that were also within our query set are listed. Displayed p-values are adjusted for multiple testing (Materials and Methods).

bioRxiv preprint doi: https://doi.org/10.1101/009464; this version posted September 21, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.


a.
MAGEC1
IL13RA1
SNORA35
GAGE10
GAGE12H
PPP1R3F
RP11-265p11.2
NAP1L6
GPR34
LINC00892
TMEM35
CENPI
ITM2A
NAP1L2
ALS_Finland
ALS_Irish
Psoriasis_CASP
Celiac_CIDR
CD_NIDDK
CD_WT1
UC_WT2
MS_case_control
MS_WT2
Vitiligo_GWAS1
Vitiligo_GWAS2
T2D_GENEVA
T2D_WT1
T1D_WT1
RA_WT1
AS_WT2
b.
MAGEC1
MCF2
CENPI
BEND2
PPP1R3F
MAGEE1
LINC00892
NLGN4X
c.
MAGEC1
RP4-562J12.2
CENPI
GAGE10
FOXP3
PPP1R3F
FANCB
NAP1L6
ARHGEF6
ITM2A
NAP1L2
d.
MIR320D2
EFHC2
MCF2
BEND2
C1GALT1C1
CENPI
NLGN4X
NDP
MAGEE2


Tissues
Whole blood
CD14+ Monocytes
CD33+ Myeloid
CD56+ NK Cells
CD4+ T cells
CD8+ T cells
BDCA4+ Dentritic Cells
CD19+ Bcells(neg. sel.)
721B lymphoblasts
CD105+ Endothelial
CD34+
Cerebellum peduncles
Cerebellum
Globus pallidus
Pons
Subthalamic nucleus
Temporal lobe
Occipital lobe
Cingulate cortex
Medulla oblongata
Parietal lobe
Caudate nucleus
Thalamus
Hypothalamus
Spinal cord
Prefrontal cortex
Amygdala
Whole brain
Pineal (day)
Pineal (night)
Retina
Skeletal muscle
Tongue
Superior cervical ganglion
Trigeminal ganglion
Skin
Atrioventricular node
Ciliary ganglion
Dorsal root ganglion
Ovary
Appendix
Uterus corpus
Heart
Liver
Colon
Small intestine
CD71+ Early Erythroid
Placenta
Lung
Prostate
Thyroid
Cardiac myocytes
Smooth muscle
Bronchial epithelial cells
Testis
Testis germ cell
Testis intersitial
Testis Leydig cell
Testis seminiferous tubule
Pancreas
Pancreatic islet
Adipocyte
Uterus
Pituitary
Salivary gland
Trachea
Olfactory bulb
Adrenal cortex
Adrenal gland
Bone marrow
Thymus
Lymph node
Tonsil
Kidney
ARHGEF6
C1GALT1C1
CENPI
EFHC2
FOXP3
IL13RA1
ITM2A
MAGEC1
MCF2
NAP1L2
NDP
NLGN4X
TMEM35
XWAS genes (all)
Genes
Expression (z-score)
7
6
5
4
3
2
1
0
-1
-2


a
ARHGEF6 (CD4+ T cells)
b
IL13RA1 (CD14+ Monocytes)
c
ITM2A (Thymus)
Tissues
Whole blood
CD14+ Monocytes
CD33+ Myeloid
CD56+ NK Cells
CD4+ T cells
CD8+ T cells
BDCA4+ DentriticCells
CD19+ Bcells(neg. sel.)
721B lymphoblasts
CD105+ Endothelial
CD34+
Cerebellum peduncles
Cerebellum
Globus pallidus
Pons
Subthalamic nucleus
Temporal lobe
Occipital lobe
Cingulate cortex
Medulla oblongata
Parietal lobe
Caudate nucleus
Thalamus
Hypothalamus
Spinal cord
Prefrontal cortex
Amygdala
Whole brain
Pineal (day)
Pineal (night)
Retina
Skeletal muscle
Tongue
Superior cervical ganglion
Trigeminal ganglion
Skin
Atrioventricular node
Ciliary ganglion
Dorsal root ganglion
Ovary
Appendix
Uterus corpus
Heart
Liver
Colon
Small intestine
CD71+ EarlyErythroid
Placenta
Lung
Prostate
Thyroid
Cardiac myocytes
Smooth muscle
Bronchial epithelial cells
Testis
Testis germ cell
Testis interstitial
Testis Leydig cell
Testis seminiferous tubule
Pancreas
Pancreatic islet
Adipocyte
Uterus
Pituitary
Salivary gland
Trachea
Olfactory bulb
Adrenal cortex
Adrenal gland
Bone marrow
Thymus
Lymph node
Tonsil
Kidney
-0.5 0 0.5 1 1.5 2 2.5 3 3.5 4
Expression (z-score)
0 1 2 3 4 5
Expression (z-score)
0 1 2 3 4 5 6 7
Expression (z-score)


CTAG1B
EXOC8
ATP5L
POLR3F
TUBGCP4
PSMB4
GTF2F1
APPBP2
COG8
MITD1
ITM2A
TYK2
MAGEC1
NFAT5
EFHC2
TNK2
ECHDC2
LRP5
TSPAN12
H2AFZ
NME1
ADAM15
NDP
BEND2
FGGY
SF3A1
CAP1
PDCD10
TGFB1I1
TCL1B
RUSC2
IL13RA1
SMAD3
HAND2
APP
DLG4
YWHAG
MCF2
SH2D1A
IL4R
BMPR1B
SMAD1
CDC42
CENPA
CST11
CENPI
IL13
DLGAP2
ATXN1
MDM2
ARHGEF7
ROCK2
RHOA
DOCK2
TGFBR2
ARHGEF6
GIT1
KAT5
PAK2
MAGEE2
IL13RA2
HDAC4
NLGN4X
SMAD2
RDX
TGFBR1
PAK1
RHOG
IL4
HDAC2
RUNX1
MON1A
PAK3
FZD4
NFATC2
WDR20
PKLR
GAGE10
GAGE12H
GPR34
NAP1L6
TRAF3IP1
HDAC9
C1GALT1
CAPN9
EPHB2
STUB1
IKZF4
ATP6V0C
HDAC7
ICOS
MAGEE1
MRPL19
GNA13
USP7
FOXP3
HSP90AA1
PARVB
CAPNS1
NFATC1
GOLGA5
IRF4
MOB1B
FANCB
NDUFV2
ZDHHC2
RORC
F3
RIBC2
DHX35
TMEM35
SP110
FANCA
FANCM
PPP1R3F
IRF1
C17orf70
LGALS3
C1GALT1C1
ZNF193
NAP1L2
B3GNT2
Associated X-linked loci
Interactors
Physical interactions
Genetic interactions
Predicted interactions