PLOS | ONE

RESEARCH ARTICLE

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

1. Department of Biological Statistics and 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.** Graduate Field of Genetics, Genomics and Development, Cornell University, Ithaca, New York, United States of America, **4.** Department of Animal and Avian Sciences, University of Maryland, College Park, Maryland, United States of America, **5.** Human Medical Genetics and Genomics Program, University of Colorado School of Medicine, Aurora, Colorado, United States of America

*ak735@cornell.edu

CrossMark
click for updates

OPEN ACCESS

Citation: Chang D, Gao F, Slavney A, Ma L, Waldman YY, et al. (2014) Accounting for eXentricities: Analysis of the X Chromosome in GWAS Reveals X-Linked Genes Implicated in Autoimmune Diseases. PLoS ONE 9(12): e113684. doi:10.1371/journal.pone.0113684

Editor: Tanja Zeller, Medical University Hamburg, University Heart Center, Germany

Received: July 14, 2014

Accepted: October 30, 2014

Published: December 5, 2014

Copyright: © 2014 Chang et al. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Data Availability: The authors confirm that all data underlying the findings are fully available without restriction. All relevant data are within the paper and its Supporting Information files.

Funding: This work was supported in part by National Institutes of Health Grant R01HG006849. AK was also supported by The Ellison Medical Foundation and the Edward Mallinckrodt, Jr. Foundation. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

Abstract

Many complex human diseases are highly sexually dimorphic, suggesting a potential contribution of the X chromosome to disease risk. However, the X chromosome has been neglected or incorrectly analyzed 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 and related diseases (AID). We associated several X-linked genes with disease risk, among which (1) *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. (2) *CENPI* is associated with three different AID, which is compelling in light of known associations with AID of autosomal genes encoding centromere proteins, as well as established autosomal evidence of pleiotropy between autoimmune diseases. (3) We replicated a previous association of *FOXP3*, a transcription factor that regulates T-cell development and function, with vitiligo; and (4) we 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, participate in major immune pathways, and display differential gene expression between males and females. Combined, the results demonstrate the importance of the X chromosome in autoimmunity, reveal

the potential of extensive XWAS, even based on existing data, and provide the tools and incentive to properly include the X chromosome in future studies.

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 it from analysis 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%) significant associations reported by GWAS of nearly 300 traits [1, 5, 6]. This 0.5% of associated SNPs is less often in functional loci compared to autosomal associated SNPs [1, 5, 7], which further suggests that X-linked associations might include a higher proportion of false positives. This is possibly due to most studies analyzing X using 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, many of 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]. Characterizing the role of X in complex diseases can provide insights into etiological differences between males and females, as well as a unique biological perspective on disease etiology because X carries a set of genes with unique functions [37–39].

X-specific considerations that are important to account for in GWAS include, but are not limited to: (1) correlation between X-linked genotype calling error rate and the sex composition of an assay 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 (due to males carrying one allele and X-inactivation in females), reduced diversity on X and other unique population genetic patterns [41–47], and a lower density of X-linked SNPs on genotyping arrays; (4) quality control (QC) criteria need to 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 lead to false positives [48–50]) between X and the autosomes; and (6) application of association tests designed for the autosomes potentially leading to statistical inaccuracies. Recent advances of association test statistics for X have been made [51–57], with a recent study discovering X-linked loci associated with height and fasting insulin level [56].

Autoimmune diseases (AID) are promising case studies for investigating the role of X in disease because they are commonly sexually dimorphic in symptoms, prevalence (most have higher prevalence in females) [10–12, 58], 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 these sexually dimorphic characteristics, a role for X-linked genes has also been suggested [27, 62, 64–66]. AID have been extensively studied by GWAS, where the majority of autosomal loci discovered have a small effect size, and the combined effect of all associated loci only explains a fraction of heritable variation in disease risk [67–69]. In addition, few of these GWAS have studied the contribution of X and, combined, have provided little evidence for its role in determining disease susceptibility [1, 5, 6].

In this study, we first introduce X-specific analytical methods and software for carrying out X-wide association studies (XWAS), which take into account several of the above 'eXentricities'. These methods apply X-specific strategies for QC, imputation, association tests, and tests of sex-specific effects. Furthermore, motivated by the unique characteristics of genes on X, we implemented the first gene-based test for associating X-linked genes and conducted an extensive XWAS of a number of AID and other diseases with a potential autoimmune component [70, 71]. Our discovery of X-linked risk genes illustrates the importance of X in AID etiology, shows that X-based analysis can be used to fruitfully mine existing datasets, and provides suitable tools and incentive for conducting such analyses. Additional XWAS can further elucidate the role of sex chromosomes in disease etiology and in the sexual dimorphism of complex diseases, which, in turn, will contribute to improved sex-specific diagnosis and treatment.

Results and Discussion

Datasets and analysis pipeline

We assembled for analysis 16 datasets of AID and other diseases (Table 1). To facilitate independent analysis and replication, we removed individuals from some datasets such that no overlapping data remained 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 the X chromosome based on whole-genome and whole-exome haplotype data from the 1000 Genomes Project (Materials and Methods). Of the 16 datasets, none of

Table 1. GWAS datasets.

Dataset	Disease	# SNPs	# Genes (#SNPs in genic regions)	# Cases (males, females)	# Controls (males, females)
ALS Finland [123]	Amyotrophic Lateral Sclerosis (ALS)	207,947	970 (72,219)	400 (198, 202)	490 (103, 387)
ALS Irish [124]	Amyotrophic Lateral Sclerosis (ALS)	219,300	967 (77,043)	221 (119, 102)	210 (112, 98)
Psoriasis CASP [128]	Psoriasis	184,246	953 (62,106)	1,209 (588, 621)	1,271 (585, 686)
Celiac Disease CIDR [125]	Celiac Disease	187,284	962 (64,836)	1,576 (447, 1129)	504 (225, 279)
CD NIDDK [127]	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* [131]	Ulcerative Colitis (UC)	196,781	963 (67,422)	2,341 (1148, 1193)	1,699 (843, 856)
MS case control [96]	Multiple Sclerosis (MS)	183,954	842 (61,119)	943 (312, 631)	851 (290, 561)
MS WT2* [132]	Multiple Sclerosis (MS)	169,707	962 (58,463)	2,666 (698, 1968)	1389 (700, 689)
Vitiligo GWAS1 [126]	Vitiligo	157,676	958 (54,384)	1,391 (411, 980)	4,521 (1985, 2536)
Vitiligo GWAS2 [133]	Vitiligo	187,688	962 (64,974)	415 (144, 271)	2,552 (973, 1579)
T2D GENEVA [129]	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)
AS WT2* [130]	Ankylosing Spondylitis (AS)	200,042	966 (69,441)	1,472 (976, 496)	1,260 (665, 595)

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, the number of genes tested in gene-based tests (# *Genes*), and the combined number of SNPs mapped to these genes or to within 15 kb of them (# *SNPs in genic regions*). 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 control subsets for each of the datasets (Materials and Methods). The size of these subsets is indicated under # *Controls*.

doi:10.1371/journal.pone.0113684.t001

the original GWAS had imputed variants in an X-specific manner, and only the Wellcome Trust Case Control Consortium 1 (WT1) carried out an analysis of X that is not identical to that of the autosomes [72].

In each of the datasets, we applied three statistical tests for association of each SNP with disease risk: FM_{02}, $FM_{F.comb}$, and $FM_{S.comb}$ (Materials and Methods). The FM_{02} test utilizes logistic regression as commonly applied in GWAS, where X-inactivation is accounted for by considering hemizygous males as equivalent to female homozygotes. The other two tests employ regression analyses separately for each sex and combine them into a single test of association using either Fisher's method ($FM_{F.comb}$) or Stouffer's method ($FM_{S.comb}$). The $FM_{F.comb}$ test accommodates the possibility of differential effect size and direction between males and females and is not affected by the allele coding in males (i.e. whether each allele in males is counted twice as in FM_{02} or only once; Materials and Methods). $FM_{S.comb}$ takes in account the potentially different sample sizes of males and females and the direction of effect, thereby increasing power in some scenarios (see Text S1). We employed EIGENSOFT [48] to remove individuals of non-European descent and to correct for potential population stratification.

Following this correction, QQ (quantile-quantile) plots for each of the three tests across all SNPs, along with genomic inflation factors, revealed no systematic bias across the datasets (Figure S1; Table S1). We provide results for association of individual SNPs with disease risk in Text S1, Figure S2, and Table S2, and focus on the results of gene-based tests (described below) for the remainder of our analysis.

We applied a gene-based test to X-linked genes in each of the 16 datasets using the FM_{02}, $FM_{F.comb}$ and $FM_{S.comb}$ statistics. Gene-based tests aggregate association signals across a group of SNPs within a locus while considering the dependence between signals due to linkage disequilibrium (LD) to assign a level of significance for the association of the locus overall. It thereby also reduces the multiple hypothesis-testing burden from the number of SNPs to the number of tested loci [73–76]. This approach can increase power for the autosomes [75, 77] and enable replication based on a different set of SNPs in the associated locus. Due to some issues discussed above (see Introduction), this increase in power can be even more pronounced for X.

For our gene-based tests, we defined genes by unique transcripts and included a flanking 15 kilobase (kb) window on each side of the transcribed region to also consider cis-regulatory elements. We used the truncated tail strength [78] and truncated product [79] methods (Materials and Methods) to combine signals across all SNPs in a gene, while accounting for LD. These two methods combine signals from several of the most significant SNPs, thus improving statistical power compared to gene-based tests that consider all SNPs or only the SNP with the strongest signal in the gene. This is especially important for cases in which a gene contains multiple risk alleles or when the causal SNP is partially tagged by multiple tested SNPs [80, 81]. From the first round of discovery, we considered for replication genes with a significance of $P<10^{-3}$ (Tables S3, S4). For these, we first attempted replication in a different dataset of the same disease (including the related Crohn's disease and ulcerative colitis), if such a dataset was available for analysis (Table 1), and applied Bonferroni correction for the number of genes we attempted to replicate. Otherwise, motivated by the shared pathogenicity of different AID [82–85] (which is also supported by our following results), we attempted replication in all other datasets considered herein (Table 1). In both cases, we attempted replication using the same test statistic that passed the first round of discovery.

Associations of X-linked genes with autoimmune and other complex diseases

We detected 54 unique genes that passed the initial discovery criterion 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 S3, S4), with overlap between the three tests due to their statistical dependence. For 42 of these 54 genes, we had an independent dataset for the same or related disease with which to attempt replication. Of these 42 genes, 5 (12%) successfully replicated,

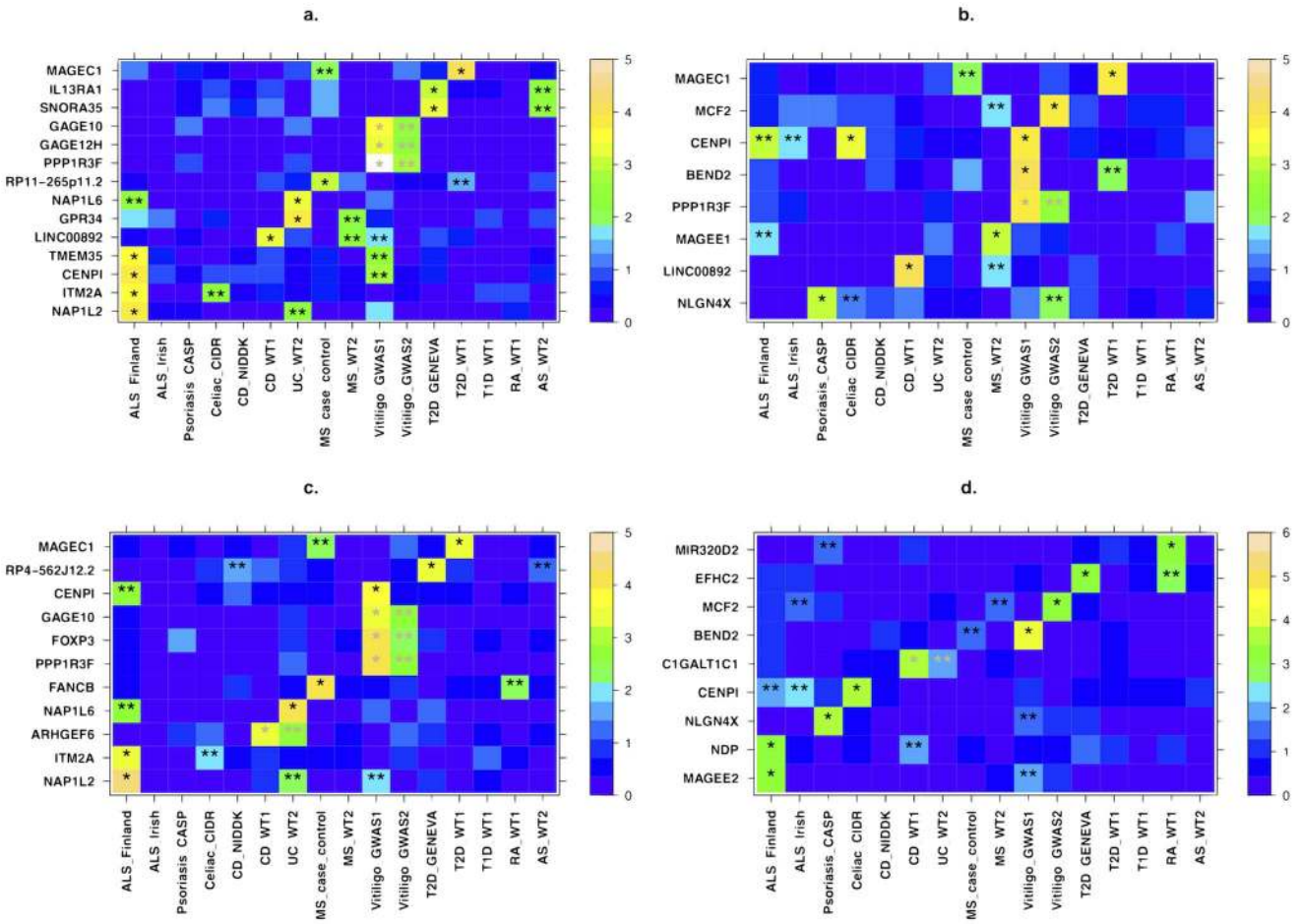

Figure 1. X-linked genes associated with autoimmune disease risk. All genes that showed evidence of association in a gene-based test and replication, including suggestive replication in any other dataset (see main text) are presented for the a) $FM_{S.comb}$ b) $FM_{F.comb}$ c) FM_{02} and d) sex-differentiated effect size tests (Materials and Methods). *X-axis* denotes the different datasets, with their names following the notation from Table 1. *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 figure are presented in Tables 2–3.

doi:10.1371/journal.pone.0113684.g001

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. To reduce the level of LD, we repeated the gene-based testing without the flanking region of 15 kb around each gene. All genes still successfully replicated in this case, though it remains unclear whether these represent independent signals or remain in LD with the same—likely unobserved—causal variant(s).

Of the above four genes we associated to vitiligo risk, *FOXP3* (combined $P=9.5\times10^{-6}$; Table 2) has been previously associated with vitiligo in a candidate gene study of this same dataset [86]. Vitiligo is a common autoimmune disorder

Table 2. Gene-based associations replicating in similar diseases.

Discovery dataset	Gene	p-value (tail, product)	Replication dataset	p-value (tail, product)	combined p-value (tail, product)
FM_{02}					
Vitiligo GWAS1	PPP1R3F	6.60×10^{-5}, 1.39×10^{-4}	Vitiligo GWAS2	8.10×10^{-3}, 2.70×10^{-3}	8.26×10^{-6}, 5.93×10^{-6}
Vitiligo GWAS1	FOXP3	1.11×10^{-4}, 2.76×10^{-4}	Vitiligo GWAS2	5.60×10^{-3}, 5.40×10^{-3}	9.50×10^{-6}, 2.15×10^{-5}
Vitiligo GWAS1	GAGE10	1.60×10^{-3}, 4.03×10^{-4}	Vitiligo GWAS2	2.80×10^{-3}, 3.80×10^{-3}	5.97×10^{-5}, 2.20×10^{-5}
CD WT1	ARHGEF6	1.70×10^{-3}, 3.66×10^{-4}	UC WT2	2.30×10^{-3}, 3.10×10^{-3}	5.26×10^{-5}, 1.67×10^{-5}
$FM_{F.comb}$					
Vitiligo GWAS1	PPP1R3F	1.14×10^{-4}, 4.96×10^{-4}	Vitiligo GWAS2	3.70×10^{-3}, 5.80×10^{-3}	6.61×10^{-6}, 3.96×10^{-5}
$FM_{S.comb}$					
Vitiligo GWAS1	PPP1R3F	6.0×10^{-6}, 7.60×10^{-5}	Vitiligo GWAS2	4.80×10^{-3}, 1.30×10^{-3}	5.29×10^{-7}, 1.69×10^{-6}
Vitiligo GWAS1	GAGE12H	6.34×10^{-4}, 6.34×10^{-4}	Vitiligo GWAS2	4.60×10^{-3}, 4.60×10^{-3}	4.01×10^{-5}, 4.01×10^{-5}
Vitiligo GWAS1	GAGE10	1.85×10^{-3}, 2.66×10^{-4}	Vitiligo GWAS2	2.90×10^{-3}, 2.80×10^{-3}	7.05×10^{-5}, 1.13×10^{-5}
Sex Difference					
CD WT1	C1GALT1C1	1.97×10^{-3}, 2.63×10^{-4}	UC WT2	1.39×10^{-2}, 1.14×10^{-2}	3.15×10^{-4}, 4.11×10^{-5}

All genes with a discovery nominal $P<1\times10^{-3}$ (in *Discovery dataset*) that also replicated in a dataset of the same or similar disease (*Replicated dataset*). Results are presented for each of the 3 tests of association, as well as for the test of sex-differential effect size, as indicated by titles in the table. For both discovery and replication, p-values of both methods of gene-based testing (truncated tail strength and truncated product) are presented. Combined p-values (last column) were calculated using Fisher's method.

doi:10.1371/journal.pone.0113684.t002

that is manifested in patches of depigmented skin due to abnormal destruction of 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 [87, 88]. Defects in the gene are also a known cause for an X-linked Mendelian autoimmunity-immunodeficiency syndrome (IPEX - immunodysregulation polyendocrinopathy enteropathy X-linked syndrome) [89].

In Crohn's Disease (CD), an inflammatory bowel disorder (IBD) with inflammation in the ileum and some regions of the colon, we discovered association of the gene *ARHGEF6* and further replicated it in the Wellcome Trust Case Control Consortium 2 (WT2) dataset for ulcerative colitis, another IBD (combined $P=1.67\times10^{-5}$). ARHGEF6 binds to a major surface protein of *H. pylori* [90], a gastric bacterium that may play a role in IBD pathology [91, 92].

We discovered that another gene, *CENPI*, was associated with three diseases (celiac disease, vitiligo, and amyotrophic lateral sclerosis (ALS)), with an overall combined $P=2.1\times10^{-7}$ (Table S5). The association of *CENPI* remains significant when combining across all 16 datasets and applying a conservative Bonferroni

correction for the number of genes we tested ($P=2.7\times10^{-5}$). *CENPI* encodes a member of a protein complex that generates spindle assembly checkpoint signals required for cell progression through mitosis [93]. CENPI is targeted by the immune system in some patients with scleroderma [94]. Additionally, autosomal genes in the same family of genes encoding centromere proteins have been previously associated with ALS (*CENPV*) [95] and with multiple sclerosis (*CENPC1*) [96]. These findings combined suggest a potential pleiotropic role of *CENPI* in risk of AID.

Motivated by the association of *CENPI* in multiple diseases, as well as previous evidence from the autosomes of shared pathogenicity across different AID [82, 83], we next sought to replicate the 54 genes from the discovery stage in diseases other than those in which they were discovered. We successfully replicated 17 genes, beyond the aforementioned 5 that replicated in the same or related disease, for a total of 22 (41%) of the 54 genes (Figure 1a–c and Table 3). Six of these 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 suggestive evidence of these genes playing a general role in autoimmunity or immune response, and we consider these genes together with the initial 5 in subsequent analyses.

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

If X-linked genes contribute to sexual dimorphism in complex diseases, then we would expect some genes to have significantly different effect sizes between males and females. We implemented a test of sex-differential effect size (Materials and Methods) and applied it across all SNPs and datasets (Materials and Methods). Consideration of QQ plots and genomic inflation factors revealed no systematic bias (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 assuming such differential effects to be of a similar nature across SNPs. We followed the same discovery and replication criteria as for the previous analyses, with detailed results provided in Figure 1d, Tables 2–3, and Table S6. Specifically, we discovered and replicated *C1GALT1C1* as exhibiting sex-differentiated effect size in risk of IBD (combined $P=4.11\times10^{-5}$). C1GALT1C1 (also known as Cosmc) is necessary for the synthesis of many O-glycan proteins [97], which are components of several antigens. Defects of C1GALT1C1 may cause Tn Syndrome, a hematological disorder [98]. We also considered replication of sex-differentiated effects in diseases other than the disease of the discovery dataset. This analysis found 8 additional genes, including both *CENPI* (combined $P=1.6\times10^{-8}$) and *MCF2* (combined $P=2.0\times10^{-4}$), which we associated with risk of AID in the above analyses (Tables 2–3). The evidence of sex-differentiated effect of each of these genes is in the same diseases as in the

Table 3. Gene-based associations replicating in other diseases.

Discovery dataset	Gene	p-value (tail, product)	Replication dataset	p-value (tail, product)	combined p-value (tail, product)
FM_{02}					
ALS Finland	NAP1L2	4.51×10^{-4}, 3.80×10^{-5}	UC WT2	5.70×10^{-3}, 3.70×10^{-3}	3.57×10^{-5}, 2.36×10^{-6}
			Vitiligo GWAS1	1.0×10^{-2}, 1.40×10^{-2}	6.00×10^{-5}, 8.22×10^{-6}
ALS Finland	ITM2A	2.10×10^{-3}, 4.10×10^{-4}	Celiac Disease CIDR	7.90×10^{-3}, 1.06×10^{-2}	1.99×10^{-4}, 5.80×10^{-5}
MS case control	FANCB	5.20×10^{-5}, 1.30×10^{-3}	RA WT1	3.80×10^{-3}, 1.10×10^{-2}	3.25×10^{-6}, 1.74×10^{-4}
Vitiligo GWAS1	CENPI	2.17×10^{-4}, 1.00×10^{-3}	ALS Finland	2.40×10^{-3}, 2.00×10^{-3}	8.06×10^{-6}, 2.82×10^{-5}
T2D GENEVA	RP4-562J12.2	4.89×10^{-4}, 1.30×10^{-4}	CD NIDDK	3.41×10^{-2}, 3.93×10^{-2}	2.00×10^{-4}, 5.56×10^{-4}
			WT2 AS	5.60×10^{-2}, 4.30×10^{-2}	3.15×10^{-4}, 7.32×10^{-5}
T2D WT1	MAGEC1	2.64×10^{-2}, 5.34×10^{-4}	MS case control	6.70×10^{-3}, 8.50×10^{-3}	1.71×10^{-3}, 6.04×10^{-5}
UC WT21	NAP1L6	1.06×10^{-3}, 5.70×10^{-5}	ALS Finland	3.10×10^{-3}, 5.50×10^{-3}	4.49×10^{-5}, 5.01×10^{-6}
$FM_{F.comb}$					
CASP	NLGN4X	8.87×10^{-4}, 1.66×10^{-2}	Vitiligo GWAS2	1.21×10^{-2}, 1.31×10^{-2}	1.34×10^{-4}, 2.05×10^{-3}
			CIDR Celiac Disease	5.10×10^{-2}, 4.90×10^{-2}	4.98×10^{-4}, 6.66×10^{-3}
Celiac CIDR	CENPI	2.90×10^{-3}, 5.23×10^{-4}	ALS Finland	1.12×10^{-2}, 1.00×10^{-3}	3.68×10^{-4}, 8.09×10^{-6}
			ALS Irish	2.68×10^{-2}, 1.64×10^{-2}	8.13×10^{-4}, 1.09×10^{-4}
			Vitiligo GWAS1	1.55×10^{-4}, 2.60×10^{-3}	7.02×10^{-6}, 1.97×10^{-5}
Vitiligo GWAS1	BEND2	1.80×10^{-3}, 7.90×10^{-5}	T2D WT1	9.30×20^{-3}, 1.29×10^{-2}	2.01×10^{-4}, 1.51×10^{-5}
Vitiligo GWAS1	CENPI	1.55×10^{-4}, 2.60×10^{-3}	ALS Finland	1.12×10^{-2}, 1.00×10^{-3}	2.48×10^{-5}, 3.60×10^{-5}
			Celiac CIDR	2.90×10^{-3}, 5.23×10^{-4}	7.02×10^{-6}, 1.97×10^{-5}
Vitiligo GWAS2	MCF2	1.70×10^{-4}, 5.76×10^{-4}	MS WT2	2.31×10^{-2}, 2.50×10^{-2}	5.28×10^{-5}, 1.75×10^{-4}
CD WT1	LINC00892	1.30×10^{-3}, 8.80×10^{-5}	MS WT2	2.42×10^{-2}, 1.99×10^{-2}	3.58×10^{-4}, 2.50×10^{-5}
T2D WT1	MAGEC1	2.75×10^{-2}, 1.81×10^{-4}	MS case control	1.42×10^{-2}, 1.50×10^{-2}	3.46×10^{-3}, 3.75×10^{-5}
MS WT2	MAGEE1	7.06×10^{-4}, 2.30×10^{-3}	ALS Finland	3.23×10^{-2}, 2.36×10^{-2}	2.67×10^{-4}, 5.87×10^{-4}
$FM_{S.comb}$					
ALS Finland	NAP1L2	5.7×10^{-4}, 1.15×10^{-4}	UC WT2	8.30×10^{-3}, 7.1×10^{-3}	6.27×10^{-5}, 1.23×10^{-5}
ALS Finland	ITM2A	8.43×10^{-4}, 3.07×10^{-4}	Celiac CIDR	6.5×10^{-3}, 1.13×10^{-2}	7.19×10^{-5}, 4.71×10^{-5}
ALS Finland	CENPI	1.27×10^{-3}, 1.75×10^{-4}	Vitiligo GWAS1	1.60×10^{-3}, 5.90×10^{-3}	2.89×10^{-5}, 1.53×10^{-5}
ALS Finland	TMEM35	2.78×10^{-3}, 3.45×10^{-4}	Vitiligo GWAS1	3.80×10^{-3}, 6.20×10^{-3}	1.31×10^{-4}, 3.01×10^{-5}
CD WT1	LINC00892	1.73×10^{-3}, 5.29×10^{-4}	MS WT2	6.30×10^{-3}, 6.40×10^{-3}	1.35×10^{-4}, 4.60×10^{-5}
			Vitiligo GWAS1	2.30×10^{-2}, 2.89×10^{-2}	4.41×10^{-4}, 1.85×10^{-4}
UC WT2	GPR34	2.62×10^{-4}, 1.62×10^{-4}	MS WT2	5.60×10^{-3}, 1.10×10^{-2}	2.12×10^{-5}, 2.54×10^{-5}
UC WT2	NAP1L6	1.19×10^{-3}, 4.29×10^{-4}	ALS Finland	4.00×10^{-3}, 1.06×10^{-2}	6.31×10^{-5}, 6.05×10^{-5}
MS case control	RP11-265P11.2	3.03×10^{-3}, 8.55×10^{-4}	T2D WT1	4.42×10^{-2}, 4.68×10^{-2}	1.32×10^{-3}, 4.45×10^{-4}
T2D GENEVA	SNORA35	2.12×10^{-3}, 4.54×10^{-4}	AS WT2	2.40×10^{-3}, 6.70×10^{-3}	6.71×10^{-5}, 4.17×10^{-5}
T2D GENEVA	IL13RA1	6.35×10^{-3}, 8.59×10^{-4}	AS WT2	6.20×10^{-3}, 7.20×10^{-3}	4.39×10^{-4}, 8.04×10^{-5}
T2D WT1	MAGEC1	2.63×10^{-2}, 6.80×10^{-5}	MS case control	1.00×10^{-2}, 1.54×10^{-2}	2.43×10^{-3}, 1.55×10^{-5}
Sex difference					
ALS Finland	MAGEE2	6.5×10^{-4}, 1.94×10^{-3}	Vitiligo GWAS1	3.08×10^{-2}, 1.64×10^{-2}	2.37×10^{-4}, 3.61×10^{-4}
ALS Finland	NDP	1.41×10^{-3}, 9.34×10^{-4}	CD WT1	8.60×10^{-3}, 1.33×10^{-2}	1.49×10^{-4}, 1.53×10^{-4}
CASP	NLGN4X	2.34×10^{-4}, 1.65×10^{-2}	Vitiligo GWAS1	4.52×10^{-2}, 4.33×10^{-2}	1.32×10^{-4}, 5.89×10^{-3}
Celiac CIDR	CENPI	4.4×10^{-3}, 2.08×10^{-4}	ALS Finland	2.03×10^{-2}, 1.78×10^{-2}	9.22×10^{-4}, 5.00×10^{-5}
			ALS Irish	9.80×10^{-3}, 4.40×10^{-3}	4.88×10^{-4}, 1.36×10^{-5}
Vitiligo GWAS1	BEND2	3.99×10^{-3}, 1.28×10^{-4}	MS case control	4.60×10^{-2}, 5.20×10^{-2}	1.76×10^{-3}, 8.60×10^{-5}

Table 3. Cont.

Discovery dataset	Gene	p-value (tail, product)	Replication dataset	p-value (tail, product)	combined p-value (tail, product)
Vitiligo GWAS2	MCF2	7.00×10^{-4}, 1.93×10^{-3}	MS WT2	2.38×10^{-2}, 2.12×10^{-2}	2.00×10^{-4}, 4.54×10^{-4}
T2D GENEVA	EFHC2	6.09×10^{-4}, 1.12×10^{-3}	RA WT1	1.58×10^{-2}, 1.40×10^{-3}	1.21×10^{-4}, 2.42×10^{-5}
RA WT1	MIR320D2	8.69×10^{-3}, 5.68×10^{-4}	ALS Irish	2.39×10^{-2}, 2.64×10^{-2}	1.97×10^{-3}, 1.82×10^{-4}

All genes with a discovery nominal $P<1\times10^{-3}$ that also replicated in a dataset of a *different* disease (see main text). The table mirrors Table 2, with the only difference being whether replication is in the same disease (Table 2) or a different one (this table). Cases in which the same association is replicated in multiple datasets span several rows.

doi:10.1371/journal.pone.0113684.t003

association analysis, thereby pointing to not only a significant contribution of the gene to risk of that disease, but also to its sex-specific effect on the same disease (Figure 1d and Table 3). We again stress that such replication in other diseases is only to be considered as suggestive evidence, and that we consider for subsequent analyses these genes together with those that replicated in the same disease.

Sex-differentiated effects could be a consequence of the X-inactivation (XCI) status of the gene, where at least 25% of human X loci escape XCI to varying degrees. There is no evidence that any of the above three genes (*C1GALT1C1*, *CENPI*, and *MCF2*) escape XCI [33, 99], and all three have degenerate Y gametologs in males; i.e. either the gene has been lost from the Y chromosome (*MCF2*) or the homologous gene on the Y is a non-functional pseudogene (*C1GALT1C1* and *CENPI*). Thus, these genes are expected to show monoallelic expression in both sexes, at least in fibroblasts in which XCI status has been derived [33, 99]. Nevertheless, it is possible that these genes show female-biased expression in other tissues as a consequence of escaping XCI in a tissue-specific or disease-specific manner [100, 101]. Additionally, the sex-differential risk factor may arise from interaction with other genes and sex-specific environmental factors.

We next directly tested whether any of the X-linked genes that we associated and replicated with AID and related disorders 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 genes that we analyzed, they exhibit 2.55-fold enrichment for differential expression between males and females as compared to all genes across all chromosomes ($P=6.5\times10^{-8}$). Unsurprisingly, *XIST*, which encodes the long non-coding RNA that induces formation of the Barr body, displays the most significant difference in gene expression between males and females among all X-linked genes ($P\ll10^{-16}$). Of the genes we associated and replicated, four exhibit significant sex-differential gene expression: *ITM2A* (4.54×10^{-9}), *EFHC2* (4.86×10^{-5}), *PPP1R3F* (7.06×10^{-5}), and *BEND2* (4.17×10^{-4}) (Materials and Methods). Importantly, two of these (*EFHC2* and *BEND2*) also passed initial discovery in the above analysis of sex-differentiated effect sizes, though they were

only replicated in datasets different than the one in which they have been discovered (Figure 1d and Table 3). These results suggest that X-linked genes associated with disease risk, especially those that exhibit sex-differentiated effect sizes, are related to sex-differential expression pattern of those genes.

Association of genes with immune-related function or Y homologs

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 concurrent association of a whole gene set with each of the individual diseases (Materials and Methods). We tested two different hypotheses by considering 3 such gene sets: The first two sets include X 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. Analysis of the immune-related gene sets was motivated by the nature of the diseases. The test of the last set, on the other hand, was motivated by the evolutionary perspective that genes with functional Y homologs are more likely to be under functional constraint since their Y homologs have survived the progressive degeneration of the Y chromosome over the course of the evolution of the supercohort *Theria* [99]. 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 using each of the 3 test statistics of association, as well with type 2 diabetes risk based on the $FM_{S.comb}$ test statistic (Table 4). Similarly, the KEGG/GO set is associated with vitiligo risk in the larger of the vitiligo datasets (Table 4). The set of genes with functional Y homologs suggestively contributes to a much larger group of AID, including psoriasis, vitiligo, celiac disease, Crohn's Disease, and type 1 diabetes, with the first two of these being significant after Bonferroni correction (Table 4). See Table S7 for detailed results for all other datasets and tests.

Relationship and biological functions of genes implicated in autoimmune disease risk

We set out to explore in three analyses the biological function of our associated disease risk genes by considering all 22 protein-coding genes we discovered and replicated with any AID or other complex disease tested. First, we investigated the gene expression patterns of 13 of these genes for which we could obtain tissue-specific expression data (Materials and Methods). Three of these genes show the highest expression in cells and organs directly involved in the immune system (Figures 2–3): *ARHGEF6* is most highly expressed in T-cells, *IL13RA1* in CD14+ monocytes, and *ITM2A* in the thymus (in which T-cells develop). Three of the remaining genes, *MCF2* (associated with vitiligo), *NAP1L2* and *TMEM35* (associated with ALS), exhibit the highest expression levels in the pineal gland

(Figure 2). The pineal gland produces and secretes melatonin, which interacts with the immune system [102, 103] and has been implicated in both vitiligo and ALS [102, 104–108], as well as suggested as a possible treatment for ALS [109].

Second, we considered co-expression of these 22 associated genes across 881 individuals (Materials and Methods). We observed that 3.9% of all X gene pairs exhibit significantly-positively correlated gene expression patterns. In comparison, 8% of pairs of genes from the set of the above 22 genes exhibit significantly-positively correlated gene expression. This significantly higher fraction relative to X genes overall (Table S8; $P = 1.53 \times 10^{-3}$) suggests that genes we associated with disease risk are more likely to work in concert and perhaps interact in the same pathways or cellular networks.

Third, we built an "interactome" by considering this set of 22 protein-coding genes along with genes they interact with in either protein-protein or genetic interactions (Materials and Methods). We found that 18 of these 22 genes are included in the same interaction network (Figure 4), which further supports that they interact with each other. 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). Another enriched pathway is that underlying lupus, which is a systemic AID. While no dataset for lupus was included in our study, the interactome is potentially enriched for genes in that pathway due to pleiotropy of genes between AID. Other significantly enriched pathways include the regulation of actin cytoskeleton, which can influence the morphology and movement of T-cells, as well as the TGF-beta signaling and ECF-receptor interaction pathways, both of which can mediate apoptosis [110, 111]. Finally, the significantly enriched Wnt signaling pathway is generally involved in cell development processes, such as cell-fate determination and cell differentiation [112]. It also plays a role in T-cell and B-cell proliferation and migration, as well as modulation of antigen presenting cells such as dendritic cells [113].

Concluding remarks

In this study, we applied an X-tailored analysis pipeline to 16 different GWAS datasets (Table 1), and thereby discovered and replicated novel associations of several genes with AID risk (Figure 1, Tables 2–3). Multiple 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 and interacting genes in immune-related pathways (Table 5; Figure 4). Several of the genes we associated with disease are involved in regulation of apoptosis, which plays a role in AID [114–116], including vitiligo [117], psoriasis [118] and rheumatoid arthritis [119]. Our analyses also highlight the sex-specific nature of associated X-linked disease risk genes shedding light on the sexual dimorphism of autoimmune and immune-mediated diseases (Figure 1, Tables 2–3).

The X chromosome has received little attention in the era of GWAS, with growing attention only during the past year [1, 56, 120, 121]. Our results highlight

Table 4. Gene set associations.

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	$FM_{s.comb}$	$\mathbf{1.64 \times 10^{-4}}$

Three curated gene sets were tested for association with disease risk. Displayed are datasets for which $P<0.05$ for association with the gene set indicated in header rows (XY homologs, Panther, KEGG/GO; Materials and Methods). Bold p-values indicate significant associations after multiple testing correction. P-values are the minimum between that based on the truncated tail strength method and the one based on the truncated product method. Results for all datasets and tests are presented in Table S7.

doi:10.1371/journal.pone.0113684.t004

the contribution of X to AID risk and yield new avenues for follow-ups, including unraveling sexual dimorphism in disease etiology. 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 by mining existing GWAS datasets. Our findings thus underscore the potential for new results and the importance of re-analyzing X in over 2,000 GWAS that have been conducted to date, especially in more recent and better powered studies than the datasets we considered here. To enable such analyses by other researchers, we have made publicly available our X chromosome analysis toolset [122] (http://keinanlab.cb.bscb.cornell.edu), which is in part an extension of PLINK [55].

Materials and Methods

Datasets

We obtained 16 GWAS datasets for analysis in this study, which are summarized in Table 1. Datasets were selected to span different autoimmune diseases, including ankylosing spondylitis, celiac disease, Crohn's disease, multiple

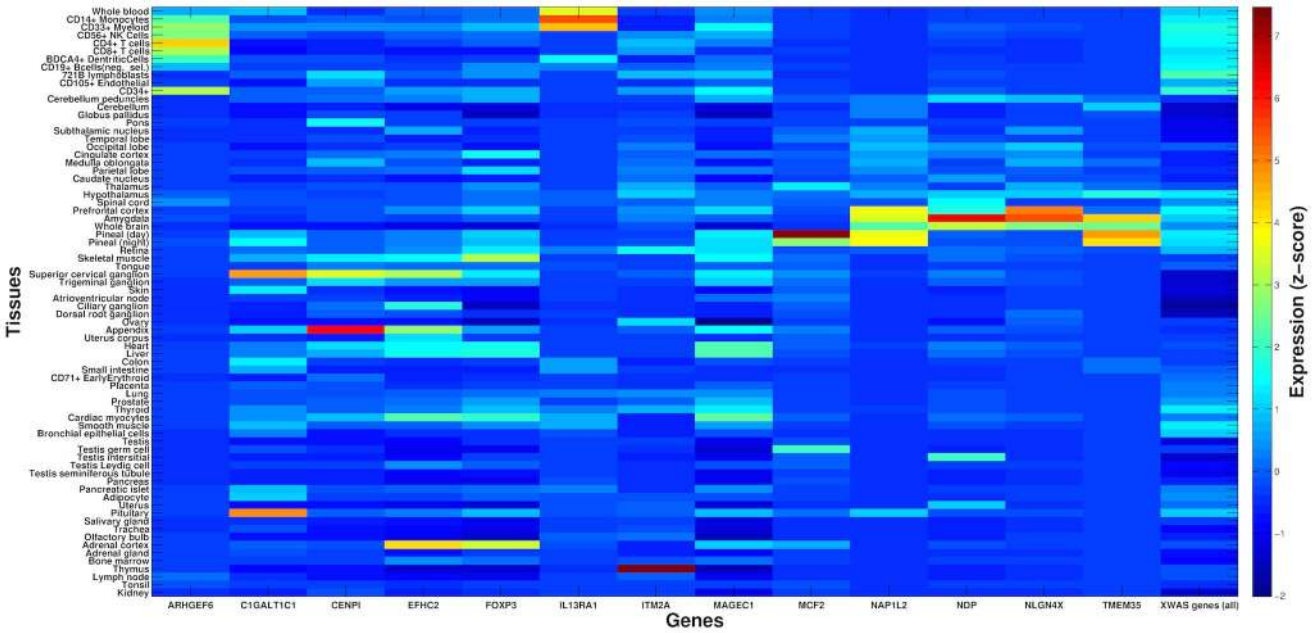

Figure 2. X-linked autoimmune disease risk genes are differentially expressed between tissues. *X-axis* presents 13 of the associated X-linked genes for which gene expression data was available for analysis. For each, a z-score is presented for the deviation of expression in each of 74 tissues (*y-axis*) 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).

doi:10.1371/journal.pone.0113684.g002

sclerosis, psoriasis, rheumatoid arthritis, type 1 diabetes, ulcerative colitis, and vitiligo. We also considered datasets of ALS and type 2 diabetes due to suggestive evidence of an autoimmune component to their etiology [70, 71].

Out of these, we obtained the following datasets from dbGaP: ALS Finland [123] (phs000344), ALS Irish [124] (phs000127), Celiac disease CIDR [125] (phs000274), MS Case Control [96] (phs000171), Vitiligo GWAS1 [126] (phs000224), CD NIDDK [127] (phs000130), CASP [128] (phs000019), and T2D GENEVA [129] (phs000091).

Additional datasets were obtained from the Wellcome Trust Case Control Consortium (WT): all WT1 [72] datasets, WT2 ankylosing spondylitis (AS) [130], WT2 ulcerative colitis (UC) [131] and WT2 multiple sclerosis (MS) [132] (Table 1). We removed overlapping control samples in order to avoid introducing any biases into replication tests. To accomplish this, we used cases from the WT1 hypertension (HT), bipolar (BP), and cardiovascular disease (CAD) datasets as additional control data. These samples were randomly distributed to the four WT1 datasets, though only BP samples were used as controls for WT1 type 2 diabetes (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.

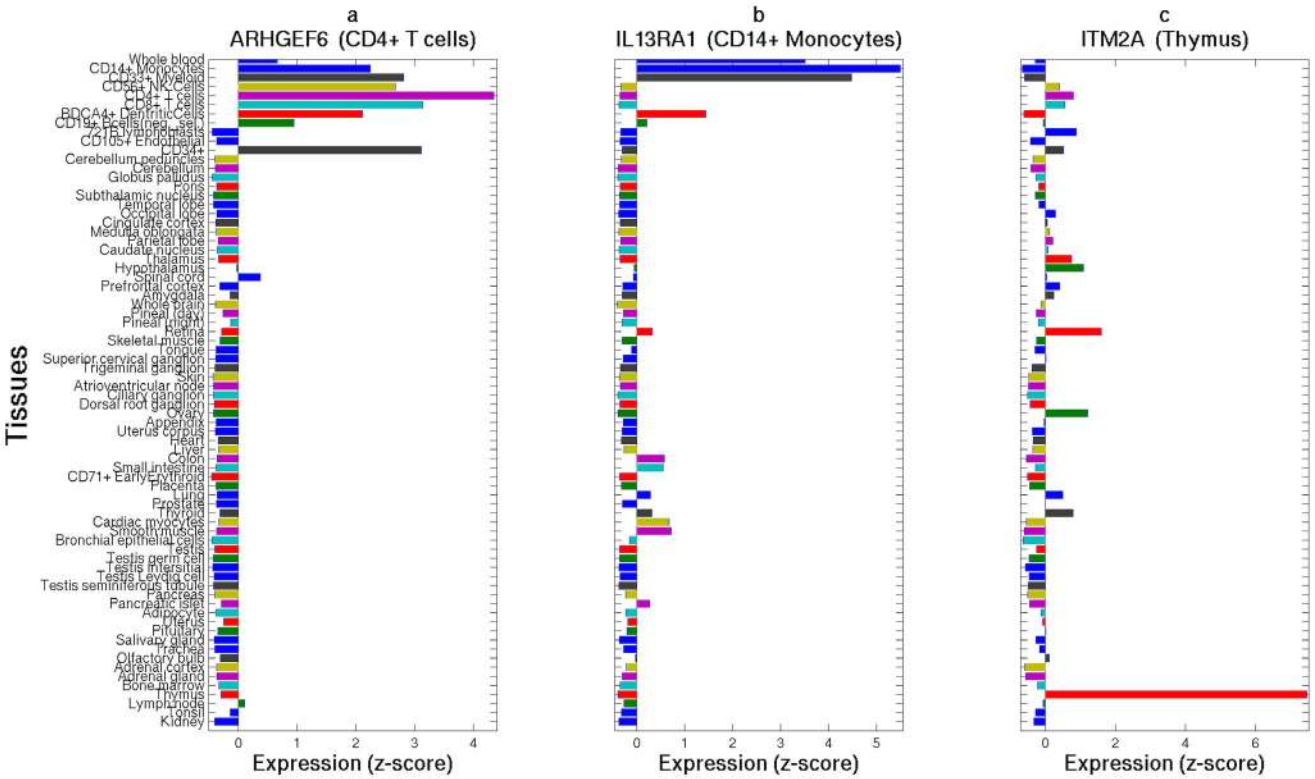

Figure 3. Three X-linked disease risk genes show high expression in immune-related tissues and cells. *ARHGEF6* (a), *IL13RA1* (b), and *ITM2A* (c) show expression greater than 4 standard deviations above the average expression of these genes in T-cells (highest in CD4+ in purple), CD14+ monocytes (blue), and the thymus (red), respectively. *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.

doi:10.1371/journal.pone.0113684.g003

We additionally analyzed the Vitiligo GWAS2 dataset [133], which similar to the Vitiligo GWAS1 dataset that we downloaded from dbGaP, contained case data only. Therefore, we obtained controls from the following datasets in dbGaP: PanScan [134, 135] (phs000206), National Institute on Aging Alzheimer's study [136] (phs000168), CIDR bone fragility [137] (phs000138), COGA [138] (phs000125), and SAGE [138–140] (phs000092). Only samples with the "general research consent" designation in these 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 [141]. We additionally filtered variants with >10%

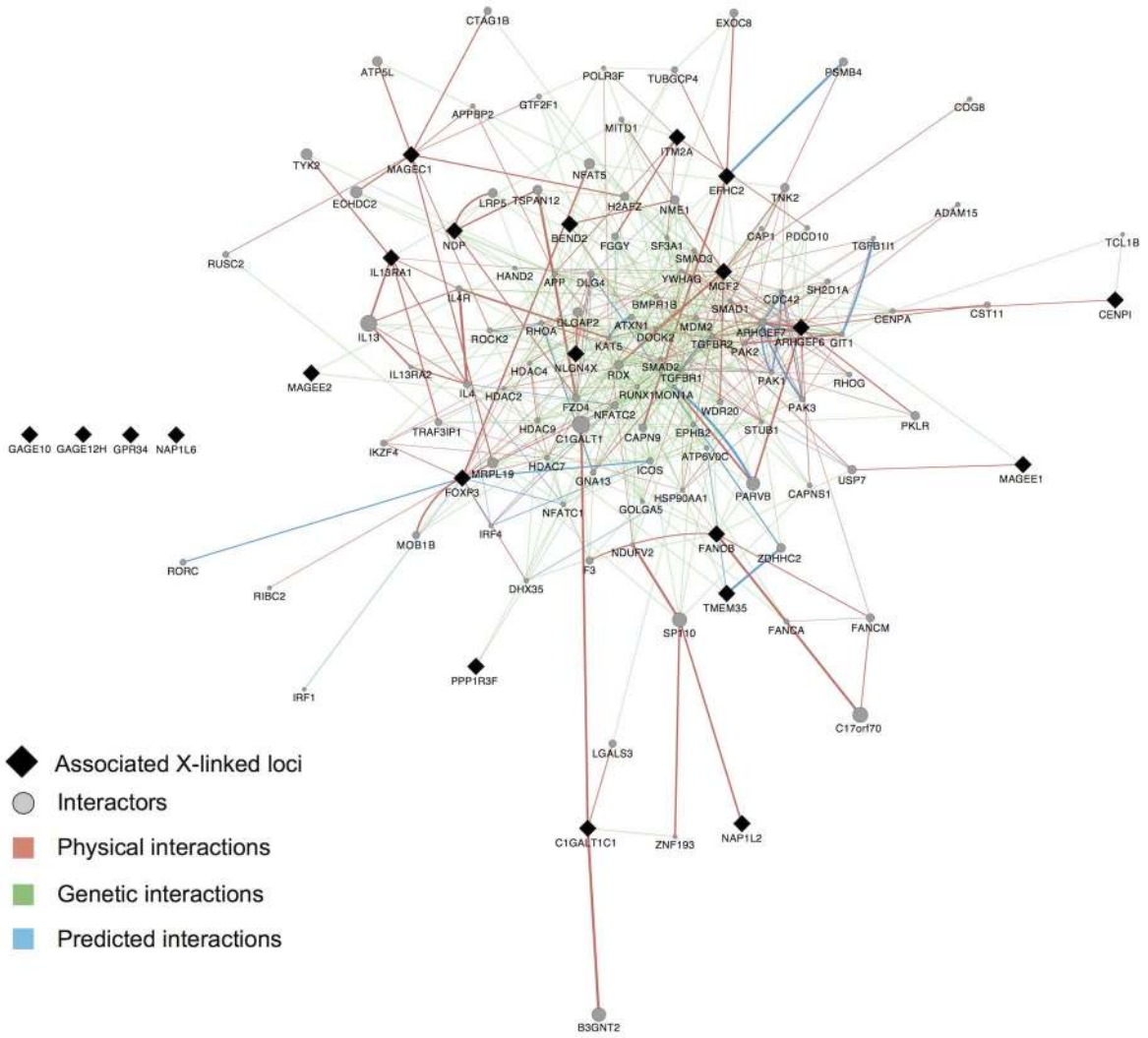

Figure 4. Interactome of X-linked disease risk genes. All 22 X-linked protein-coding genes that showed evidence of association and replication (Figure 1) are denoted by black diamonds and are presented together with genes that interact with them (grey circles) (Materials and Methods). *Physical interactions* 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.

doi:10.1371/journal.pone.0113684.g004

missingness, variants with a minor allele frequency (MAF) <0.005, and variants for which missingness was significantly correlated with phenotype ($P<1\times10^{-4}$). X-specific QC steps included filtering variants that are not in Hardy-Weinberg equilibrium in females ($P<1\times10^{-4}$) 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 [120]. However, since we observed no difference in the significant results when applying it to two of the datasets (CD NIDDK, MS case control), we considered data prior to this QC step in our

Table 5. Gene-enrichment analysis of the interactome.

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}

Genes we discovered and replicated as associated with any disease tested, and their interacting genes (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).

doi:10.1371/journal.pone.0113684.t005

analyses. Finally, following all above QC steps, we removed variants that exhibit differential missingness between males and females ($P<10^{-7}$) [120, 142, 143]. This step follows the procedure described by König et al. [120] based on a χ^2 test.

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. [42, 46, 144–147]). 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 on the X captures a 1:2 male to female contribution, while on the autosomes males and females contribute equally to the observed structure. Ideally, population structure on the X needs to be considered to accurately correct for population stratification in 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 available to infer structure 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 studies in admixed populations, consideration of population structure on the X chromosome alone can provide a more accurate population stratification correction for XWAS. For

example, though African Americans have on average ~80% African and ~20% European ancestry, they exhibit a significant deviation across the X chromosome from these genome-wide estimates. Specifically, African ancestry levels are higher on X, which is due to the sex-biased admixture in which ancestors included relatively more African females and, correspondingly, more European males [148]. Hence, population structure estimated genome-wide (e.g. by EIGENSOFT) may not accurately correct for population stratification in testing X-linked loci in studies of African Americans.

All subsequent analyses are hence based on first excluding any individuals inferred based on EIGENSOFT [48] to be of non-European ancestry. 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 original studies. Principal component covariates were not added to the regression model for the ALS Finland, ALS Irish, and CASP datasets as no inflated p-values were observed in these studies [123, 124, 128] (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 IMPUTE2 is to account for the reduced effective population size (*Ne*) of the X chromosome by assuming that it is 25% less than that of the autosomes, thereby improving imputation accuracy 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.

Single-SNP association analysis

We considered 3 tests for associating 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 our tests and results of sex-differentiated effect size and sex-differentiated 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 test 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. Implementation follows the equations as provided by Willer *et al.* [153]. Power calculations for these 3 test statistics for a few simulated examples are provided in the Text S1.

Gene-based association analysis

Based on all single-SNP association tests, we implemented an equivalent gene-based test for each statistic by considering all SNPs across each gene. This was carried out in the general framework of VEGAS [73], where the significance of an observed gene-based test statistic is assessed from the distribution of test statistics that is expected given LD between SNPs in that gene [73]. Specifically, the n observed SNP-level test statistics are summed together, where n represents the number of SNPs in a gene. Next, simulated statistics are obtained as follows: n statistics are randomly drawn from a multivariate normal (MVN) distribution and summed. The MVN distribution has a mean of 0 and an $n \times n$ covariance matrix corresponding to the pairwise LD between SNPs mapped to the gene. This procedure is repeated k times in order to obtain a distribution of gene-based statistics. The significance is then calculated as the proportion of the k simulations that produced statistics that were as or more extreme than the observed one.

Here, we have implemented a slight modification to this procedure: Instead of summing the SNP-based test statistics themselves, we combined SNP-based p-values with either the truncated tail strength [78] or the truncated product [79] method, which have been suggested to be more powerful in some scenarios [80, 81]. The simulation procedure is carried out as above, where simulated p-values are derived from the simulated test statistics. To increase time efficiency of the simulation procedure, k was determined adaptively 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 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

In a fourth test, we assayed the difference in the effect size between males and females at each SNP based on statistics derived from the sex-stratified test described above. Considering the female-only and male-only statistics, 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* stands for the odds ratio estimated in either the male-only or female-only test, *SE* is the standard error in either test, 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. Finally, we combined the single-SNP tests in each gene into a gene-based test of sex-differential effect size along the same lines as described above for the association test statistics.

Tests of sex-difference and correlation of gene expression

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. Using a hypergeometric test, we 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. For each gene, we then tested for differential expression between males and females using the Wilcoxon rank sum test across individuals and applied Bonferroni correction to its p-values. We assessed whether any of the 22 protein-coding genes that were associated and replicated in any dataset (Figure 1; Tables 2–3) showed significant sex-differential expression. 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 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 22 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 using 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 genetic, protein-protein, and predicted interactors with the highest interaction confidence scores. 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 database [162]. Enrichment was assessed with the hypergeometric test [160] and reported p-values were adjusted for multiple testing using the Benjamini-Hochberg FDR correction as suggested for such analyses [160]. Pathways that only included a single gene from our 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 different 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 S4, S5). Thus, the results we report are based on 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 considered all genes from the two databases in 15 and 14 categories, respectively, that are particularly relevant for autoimmune response. We subsequently removed eight genes from this list that we found were either too general (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 S9); (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 Wilson-Sayres & Makova [99].

Supporting Information

Figure S1. QQ-plots for single marker association tests. Blue triangles denote association p-values for the $FM_{F.comb}$ test, red crosses denote p-values for the $FM_{S.comb}$, while the black points denote association p-values for the FM_{02} test. P-values are plotted on a log scale. Respective genomic inflation factors are summarized in Table S1.
doi:10.1371/journal.pone.0113684.s001 (TIFF)

Figure S2. Significant SNP associations. (a) A Manhattan plot of the nominal p-values for the FM_{02} (upper), $FM_{F.comb}$ (middle), and $FM_{S.comb}$ (lower) tests of association for chromosome X SNPs in the 16 datasets. The dotted purple lines correspond to the X-chromosome-wide significance threshold for each dataset. The significant associations are shown as red diamonds. (b–c) Regional association plots of the association results of the FM_{02} test and LD for (b) Vitiligo GWAS1 dataset and (c) WT2 UC dataset. LD structure was plotted using a revised version of the *snp.plotter* software [165]. Due to the large number of SNPs in the associated region of Vitiligo GWAS1, only 1 in every 10 of the non-significantly associated SNPs is shown. We focus on regions presented in (b) and (c) since they show the typical LD peaks around significant association signals.
doi:10.1371/journal.pone.0113684.s002 (TIFF)

Figure S3. QQ-plots for test of sex-differentiated effect size. Similar to Figure S1, except that p-values are for the test of differential effect size between males and females. Respective genomic inflation factors are summarized in Table S1.
doi:10.1371/journal.pone.0113684.s003 (TIFF)

Figure S4. Simulation versus permutation derived p-values for gene-set tests for FM_{02}. Comparison between simulation derived (*x-axis*) and permutation derived (*y-axis*) p-values for the gene-set association analysis using the FM_{02} test statistic. *r* represents Pearson's correlation coefficient and the significance of the correlation is indicated in parentheses in scientific notation.
doi:10.1371/journal.pone.0113684.s004 (TIFF)

Figure S5. Simulation versus permutation derived p-values for gene-set tests for $FM_{F.comb}$. Similar to Figure S4 except for considering the $FM_{F.comb}$ test statistic.
doi:10.1371/journal.pone.0113684.s005 (TIFF)

Table S1. Genomic inflation factors were calculated from the observed p-values in the various tests. No inflation factor exceeds 1.14. Together with the respective QQ-plots (Figures S1 and S3) these results suggest little to no inflation in the observed SNP-level p-values.
doi:10.1371/journal.pone.0113684.s006 (DOC)

Table S2. All significant associations (adjusted P<0.05) as observed in Figure S2. P-values are Bonferroni adjusted for the number of SNPs tested (Table 1).
doi:10.1371/journal.pone.0113684.s007 (DOC)

Table S3. All genes with either truncated tail or truncated product $P<1\times10^{-3}$ for the $FM_{F.comb}$ and the $FM_{S.comb}$ tests.
doi:10.1371/journal.pone.0113684.s008 (DOC)

Table S4. All genes with either truncated tail or truncated product $P<1\times10^{-3}$ for the FM_{02} test.
doi:10.1371/journal.pone.0113684.s009 (DOC)

Table S5. CENPI association p-values for the $FM_{F.comb}$ test across the 16 datasets.
doi:10.1371/journal.pone.0113684.s010 (DOC)

Table S6. All genes with either the truncated tail or truncated product $P<1\times10^{-3}$ for the sex difference test.
doi:10.1371/journal.pone.0113684.s011 (DOC)

Table S7. All p-values for all gene sets and all datasets are listed. Those with P<0.05 are highlighted in Table 4 in the main text.
doi:10.1371/journal.pone.0113684.s012 (DOC)

Table S8. Pairs of X-linked genes that are significantly co-expressed. Presented are pairs of genes that are significantly co-expressed, after multiple hypothesis correction, along with the squared Spearman's correlation coefficient (r^2) and p-value of a Spearman's rank correlation test (Materials and Methods).
doi:10.1371/journal.pone.0113684.s013 (DOCX)

Table S9. List of genes in the KEGG/GO immune gene set.
doi:10.1371/journal.pone.0113684.s014 (DOC)

Text S1. Supplementary information detailing single-SNP association analysis and power calculations for gene-based tests.
doi:10.1371/journal.pone.0113684.s015 (DOCX)

Acknowledgments

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 also 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.

Author Contributions

Conceived and designed the experiments: AK DC AJS AS YYW LM. Performed the experiments: DC FG AJS AS YYW. Analyzed the data: DC FG AJS AS YYW. Contributed reagents/materials/analysis tools: AK RS FG AJS DC AS YYW. Wrote the paper: AK FG DC AJS AS YYW PB LM AM RS.

References

1. **Wise AL, Gyi L, Manolio TA** (2013) eXclusion: toward integrating the X chromosome in genome-wide association analyses. American Journal of Human Genetics 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 Research 33: D514–517.
3. **Amberger J, Bocchini CA, Scott AF, Hamosh A** (2009) McKusick's Online Mendelian Inheritance in Man (OMIM). Nucleic Acids Research 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)). Human Mutation 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, Morales J, Junkins HA, Hall PN, et al.** (2013) A Catalog of Published Genome-wide Association Studies. Available: http://www.genome.gov/gwastudies/. Accessed 2014 Nov 15.
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. Proceedings of the National Academy of Sciences of the United States of America 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. Nature Immunology 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. Archives of general 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. Cerebral Cortex 11: 490–497.
17. **Jazin E, Cahill L** (2010) Sex differences in molecular neuroscience: from fruit flies to humans. Nature reviews Neuroscience 11: 9–17.
18. **Choi BG, McLaughlin MA** (2007) Why men's hearts break: cardiovascular effects of sex steroids. Endocrinology and metabolism clinics of North America 36: 365–377.

19. **Anderson KM, Odell PM, Wilson PW, Kannel WB** (1991) Cardiovascular disease risk profiles. American heart journal 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. American heart journal 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 Research 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. Journal of the National Cancer Institute 88: 183–192.

27. **Ober C, Loisel Da, Gilad Y** (2008) Sex-specific genetic architecture of human disease. Nature Reviews Genetics 9: 911–922.

28. **Patsopoulos NA, Tatsioni A, Ioannidis JP** (2007) Claims of sex differences: an empirical assessment in genetic associations. JAMA : the journal of the American Medical Association 298: 880–893.

29. **Fish EN** (2008) The X-files in immunity: sex-based differences predispose immune responses. Nature reviews Immunology 8: 737–744.

30. **Nelson JL, Ostensen M** (1997) Pregnancy and rheumatoid arthritis. Rheumatic diseases clinics of North America 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. The New England Journal of Medicine 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. Nature Genetics 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. Evolutionary applications 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 research : an international journal on the molecular, supramolecular and evolutionary aspects of chromosome biology 17: 811–820.

38. **Saifi GM, Chandra HS** (1999) An apparent excess of sex- and reproduction-related genes on the human X chromosome. Proceedings Biological sciences/The Royal Society 266: 203–209.

39. **Nguyen DK, Disteche CM** (2006) High expression of the mammalian X chromosome in brain. Brain Research 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. American journal of medical genetics Part A 129A: 136–143.

41. **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. Nature Genetics 39: 1251–1255.

42. **Hammer MF, Mendez FL, Cox MP, Woerner AE, Wall JD** (2008) Sex-biased evolutionary forces shape genomic patterns of human diversity. PLoS Genetics 4: e1000202.

43. **Keinan A, Mullikin JC, Patterson N, Reich D** (2009) Accelerated genetic drift on chromosome X during the human dispersal out of Africa. Nature Genetics 41: 66–70.

44. **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. Nature Genetics 42: 830–831.

45. **Lohmueller KE, Degenhardt JD, Keinan A** (2010) Sex-averaged recombination and mutation rates on the X chromosome: a comment on Labuda et al. American Journal of Human Genetics 86: 978–980; author reply 980-971.

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

47. **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. Nature Genetics 43: 741–743.

48. **Patterson N, Price AL, Reich D** (2006) Population structure and eigenanalysis. PLoS Genetics 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. Nature Genetics 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. Genetic Epidemiology 31: 834–843.

52. **Clayton DG** (2009) Sex chromosomes and genetic association studies. Genome medicine 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. Genetic Epidemiology 36(5): 438–450.

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. American Journal of Human Genetics 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 Genetics 10: e1004127.

57. **Loley C, Ziegler A, Konig IR** (2011) Association tests for X-chromosomal markers–a comparison of different test statistics. Human heredity 71: 23–36.

58. **Gleicher N, Barad DH** (2007) Gender as risk factor for autoimmune diseases. Journal of Autoimmunity 28: 1–6.

59. **Beeson PB** (1994) Age and sex associations of 40 autoimmune diseases. The American journal of medicine 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. Proceedings of the National Academy of Sciences of the United States of America 107: 15838–15843.

62. **Selmi C, Brunetta E, Raimondo MG, Meroni PL** (2012) The X chromosome and the sex ratio of autoimmunity. Autoimmunity Reviews 11: A531–537.

63. **Tiniakou E, Costenbader KH, Kriegel MA** (2013) Sex-specific environmental influences on the development of autoimmune diseases. Clinical Immunology 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. Journal of Autoimmunity 38: J109–119.

65. **Libert C, Dejager L, Pinheiro I** (2010) The X chromosome in immune functions: when a chromosome makes the difference. Nature reviews Immunology 10: 594–604.

66. **Bianchi I, Lleo A, Gershwin ME, Invernizzi P** (2011) The X chromosome and immune associated genes. Journal of autoimmunity 38(2–3): J187–J192.

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 60(3): 189–196.

71. **Pagani MR, Gonzalez LE, Uchitel OD** (2011) Autoimmunity in amyotrophic lateral sclerosis: past and present. Neurology research international 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. American Journal of Human Genetics 87: 139–145.

74. **Neale BM, Sham PC** (2004) The future of association studies: gene-based analysis and replication. American Journal of Human Genetics 75: 353–362.

75. **Beyene J, Tritchler D, Asimit JL, Hamid JS** (2009) Gene- or region-based analysis of genome-wide association studies. Genetic Epidemiology 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. American Journal of Human Genetics 88: 283–293.

77. **Neale BM, Medland SE, Ripke S, Asherson P, Franke B, et al.** (2010) Meta-analysis of genome-wide association studies of attention-deficit/hyperactivity disorder. Journal of the American Academy of Child and Adolescent Psychiatry 49: 884–897.

78. **Jiang B, Zhang X, Zuo Y, Kang G** (2011) A powerful truncated tail strength method for testing multiple null hypotheses in one dataset. Journal of Theoretical Biology 277: 67–73.

79. **Zaykin DV, Zhivotovsky LA, Westfall PH, Weir BS** (2002) Truncated product method for combining P-values. Genetic Epidemiology 22: 170–185.

80. **Ma L, Clark AG, Keinan A** (2013) Gene-Based Testing of Interactions in Association Studies of Quantitative Traits. PLoS Genetics 9.

81. **Huang HL, Chanda P, Alonso A, Bader JS, Arking DE** (2011) Gene-Based Tests of Association. PLoS Genetics 7.

82. **Sirota M, Schaub MA, Batzoglou S, Robinson WH, Butte AJ** (2009) Autoimmune disease classification by inverse association with SNP alleles. PLoS genetics 5: e1000792.

83. **Cotsapas C, Voight BF, Rossin E, Lage K, Neale BM, et al.** (2011) Pervasive sharing of genetic effects in autoimmune disease. PLoS Genetics 7: e1002254.

84. **Sivakumaran S, Agakov F, Theodoratou E, Prendergast JG, Zgaga L, et al.** (2011) Abundant pleiotropy in human complex diseases and traits. American Journal of Human Genetics 89: 607–618.

85. **Chang D, Keinan A** (2014) Principal component analysis characterizes shared pathogenetics from genome-wide association studies. PLOS Comput Biol 10: e1003820.

86. **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. The Journal of investigative dermatology 131: 371–381.

87. **Tang QZ, Bluestone JA** (2008) The Foxp3(+) regulatory T cell: a jack of all trades, master of regulation. Nature Immunology 9: 239–244.

88. **Fontenot JD, Gavin MA, Rudensky AY** (2003) Foxp3 programs the development and function of CD4(+)CD25(+) regulatory T cells. Nature Immunology 4: 330–336.

89. **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. Nature Genetics 27: 20–21.

90. **Baek HY, Lim JW, Kim H** (2007) Interaction between the Helicobacter pylori CagA and alpha-Pix in gastric epithelial AGS cells. Annals of the New York Academy of Sciences 1096: 18–23.

91. **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. Inflammatory bowel diseases 16: 1077–1084.

92. **Jin X, Chen YP, Chen SH, Xiang Z** (2013) Association between Helicobacter Pylori infection and ulcerative colitis–a case control study from China. International journal of medical sciences 10: 1479–1484.

93. **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 & Development 26: 542–547.

94. **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. Clinica chimica acta; international journal of clinical chemistry 412: 2267–2271.

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. Neurobiology of Aging 34: 357 e357-319.

96. **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. Human Molecular Genetics 18: 767–778.

97. **Ju T, Cummings RD** (2005) Protein glycosylation: chaperone mutation in Tn syndrome. Nature 437: 1252.

98. **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. European Journal of Immunology 22: 1835–1842.

99. **Wilson Sayres MA, Makova KD** (2013) Gene survival and death on the human Y chromosome. Molecular Biology and Evolution 30: 781–787.

100. **Sharp A, Robinson D, Jacobs P** (2000) Age- and tissue-specific variation of X chromosome inactivation ratios in normal women. Human Genetics 107: 343–349.

101. **Cotton AM, Lam L, Affleck JG, Wilson IM, Penaherrera MS, et al.** (2011) Chromosome-wide DNA methylation analysis predicts human tissue-specific X inactivation. Human Genetics 130: 187–201.

102. **Calvo JR, Gonzalez-Yanes C, Maldonado MD** (2013) The role of melatonin in the cells of the innate immunity: a review. Journal of Pineal Research 55: 103–120.

103. **Pohanka M** (2013) Impact of melatonin on immunity: a review. Central European Journal of Medicine 8: 369–376.

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

105. **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. Journal of Pineal Research 33: 186–187.

106. **Terry PD, Villinger F, Bubenik GA, Sitaraman SV** (2009) Melatonin and Ulcerative Colitis: Evidence, Biological Mechanisms, and Future Research. Inflammatory bowel diseases 15: 134–140.

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

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

109. **Weishaupt JH, Bartels C, Polking E, Dietrich J, Rohde G, et al.** (2006) Reduced oxidative damage in ALS by high-dose enteral melatonin treatment. Journal of Pineal Research 41: 313–323.

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

111. **Lukashev ME, Werb Z** (1998) ECM signalling: orchestrating cell behaviour and misbehaviour. Trends in Cell Biology 8: 437–441.

112. **Logan CY, Nusse R** (2004) The Wnt signaling pathway in development and disease. Annual Review of Cell and Developmental Biology 20: 781–810.

113. **Staal FJ, Luis TC, Tiemessen MM** (2008) WNT signalling in the immune system: WNT is spreading its wings. Nature reviews Immunology 8: 581–593.

114. **Eguchi K** (2001) Apoptosis in autoimmune diseases. Internal medicine 40: 275–284.

115. **Kawakami A, Eguchi K** (2002) Involvement of apoptotic cell death in autoimmune diseases. Medical electron microscopy : official journal of the Clinical Electron Microscopy Society of Japan 35: 1–8.

116. **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. Proceedings of the National Academy of Sciences of the United States of America 110: 2599–2604.

117. **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. Histology and histopathology 24: 849–857.

118. **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. The Journal of investigative dermatology 131: 1916–1926.

119. **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. European review for medical and pharmacological sciences 18: 78–85.

120. **Konig IR, Loley C, Erdmann J, Ziegler A** (2014) How to include chromosome x in your genome-wide association study. Genetic Epidemiology 38: 97–103.

121. **Conde L, Foo JN, Riby J, Liu J, Darabi H, et al.** (2013) X chromosome-wide association study of follicular lymphoma. British Journal of Haematology 162: 858–862.

122. **Chang D, Gao F, Keinan A** (2014) XWAS: a toolset for genetic data analysis and association studies of the X chromosome. bioRxiv. doi: http://dx.doi.org/10.1101/009795.

123. **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 neurology 9: 978–985.

124. **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. Human Molecular Genetics 17: 768–774.

125. **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.

126. **Jin Y, Birlea SA, Fain PR, Gowan K, Riccardi SL, et al.** (2010) Variant of TYR and autoimmunity susceptibility loci in generalized vitiligo. The New England Journal of Medicine 362: 1686–1697.

127. **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.

128. **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. Nature Genetics 41: 199–204.

129. **Qi L, Cornelis MC, Kraft P, Stanya KJ, Kao WHL, et al.** (2010) Genetic variants at 2q24 are associated with susceptibility to type 2 diabetes. Human Molecular Genetics 19: 2706–2715.

130. **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. Nature Genetics 43: 761–767.

131. **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. Nature Genetics 41: 1330–1334.

132. **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.

133. **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. Nature Genetics 44: 676–680.

134. **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. Nature Genetics 42: 224–228.

135. **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. Nature Genetics 41: 986–990.

136. **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. Archives of Neurology 65: 1518–1526.

137. **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. Nature Genetics 44: 491–501.

138. **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 research & health : the journal of the National Institute on Alcohol Abuse and Alcoholism 26: 208–213.

139. **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 and Alcohol Dependence 95: 14–22.

140. **Bierut LJ** (2007) Genetic variation that contributes to nicotine dependence. Pharmacogenomics 8: 881–883.

141. **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. Genetic Epidemiology 34: 591–602.

142. **Ling H, Hetrick K, Bailey-Wilson JE, Pugh EW** (2009) Application of sex-specific single-nucleotide polymorphism filters in genome-wide association data. BMC proceedings 3 Suppl 7: S57.

143. **Ziegler A** (2009) Genome-wide association studies: quality control and population-based measures. Genetic Epidemiology 33 Suppl 1: S45–50.

144. **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. Nature Genetics 29: 20–21.

145. **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. Nature Genetics 36: 1122–1125.

146. **Heyer E, Chaix R, Pavard S, Austerlitz F** (2012) Sex-specific demographic behaviours that shape human genomic variation. Molecular Ecology 21: 597–612.

147. **Arbiza L, Gottipati S, Siepel A, Keinan A** (2014) Contrasting X-linked and autosomal diversity across 14 human populations. Am J Hum Genet 94: 827–844.

148. **Bryc K, Auton A, Nelson MR, Oksenberg JR, Hauser SL, et al.** (2010) Genome-wide patterns of population structure and admixture in West Africans and African Americans. Proc Natl Acad Sci U S A 107: 786–791.

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 Genetics 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 Genetics 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. European Journal of Epidemiology 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 Research 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. Proceedings of the National Academy of Sciences of the United States of America 101: 6062–6067.

158. **Wu C, Macleod I, Su AI** (2013) BioGPS and MyGene.info: organizing online, gene-centric information. Nucleic Acids Research 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 Research 38: W214–220.

160. **Wang J, Duncan D, Shi Z, Zhang B** (2013) WEB-based GEne SeT AnaLysis Toolkit (WebGestalt): update 2013. Nucleic Acids Research. doi: 10.1093/nar/gkt439.

161. **Zhang B, Kirov S, Snoddy J** (2005) WebGestalt: an integrated system for exploring gene sets in various biological contexts. Nucleic Acids Research 33: W741–748.

162. **Kanehisa M, Goto S** (2000) KEGG: kyoto encyclopedia of genes and genomes. Nucleic Acids Research 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. Nature Genetics 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 Research 13: 2129–2141.

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