nature genetics

A genome-wide association study in Han Chinese identifies a susceptibility locus for primary Sjögren's syndrome at 7q11.23

Yongzhe Li[1,42], Kunlin Zhang[2,42], Hua Chen[1,42], Fei Sun[1], Juanjuan Xu[1], Ziyan Wu[1], Ping Li[1], Liuyan Zhang[2], Yang Du[2], Haixia Luan[1], Xi Li[1], Lijun Wu[3], Hongbin Li[4], Huaxiang Wu[5], Xiangpei Li[6], Xiaomei Li[6], Xiao Zhang[7], Lu Gong[8], Lie Dai[9], Lingyun Sun[10], Xiaoxia Zuo[11], Jianhua Xu[12], Huiping Gong[13], Zhijun Li[14], Shengquan Tong[15], Min Wu[16], Xiaofeng Li[17], Weiguo Xiao[18], Guochun Wang[19], Ping Zhu[20], Min Shen[1], Shengyun Liu[21], Dongbao Zhao[22], Wei Liu[23], Yi Wang[24], Cibo Huang[25], Quan Jiang[26], Guijian Liu[27], Bin Liu[28], Shaoxian Hu[29], Wen Zhang[1], Zhuoli Zhang[30], Xin You[1], Mengtao Li[1], Weixin Hao[31], Cheng Zhao[32], Xiaomei Leng[1], Liqi Bi[33], Yongfu Wang[34], Fengxiao Zhang[35], Qun Shi[1], Wencheng Qi[36], Xuewu Zhang[37], Yuan Jia[37], Jinmei Su[1], Qin Li[38], Yong Hou[1], Qingjun Wu[1], Dong Xu[1], Wenjie Zheng[1], Miaojia Zhang[39], Qian Wang[1], Yunyun Fei[1], Xuan Zhang[1], Jing Li[1], Ying Jiang[1], Xinping Tian[1], Lidan Zhao[1], Li Wang[1], Bin Zhou[40], Yang Li[41], Yan Zhao[1], Xiaofeng Zeng[1], Jurg Ott[2], Jing Wang[2,43] & Fengchun Zhang[1,43]

Primary Sjögren's syndrome is one of the most common autoimmune diseases. So far, genetic studies of Sjögren's syndrome have relied mostly on candidate gene approaches. To identify new genetic susceptibility loci for primary Sjögren's syndrome, we performed a three-stage genome-wide association study in Han Chinese. In the discovery stage, we analyzed 556,134 autosomal SNPs in 542 cases and 1,050 controls. We then validated promising associations in 2 replication stages comprising 1,303 cases and 2,727 controls. The combined analysis identified *GTF2I* at 7q11.23 (rs117026326: $P_{combined}$ = 1.31 × 10^{-53}, combined odds ratio ($OR_{combined}$) = 2.20) as a new susceptibility locus for primary Sjögren's syndrome. Our analysis also confirmed previously reported associations in Europeans in the regions of *STAT4*, *TNFAIP3* and the major histocompatibility complex (MHC). Fine mapping of the region around *GTF2I* showed that rs117026326 in *GTF2I* had the most significant association, with associated SNPs extending from *GTF2I* to *GTF2IRD1-GTF2I*.

Sjögren's syndrome is a systemic autoimmune disease characterized by lymphocytic infiltration of exocrine glands and epithelia at multiple sites. Affected individuals exhibit a persistent feeling of dry mouth (xerostomia) and dry eyes (keratoconjunctivitis sicca). Sjögren's syndrome may occur alone (defined as primary Sjögren's syndrome) or in association with another defined autoimmune disease such as rheumatoid arthritis or systemic lupus erythematosus (SLE) (defined as secondary Sjögren's syndrome)[1]. Primary Sjögren's syndrome is one of the most common autoimmune diseases. The prevalence of primary Sjögren's syndrome is estimated to be 0.05–4.8% globally[2] and 0.77% in China[3]. At present, genetic studies of primary Sjögren's syndrome have mostly relied on candidate gene approaches[4], with immune-related genes as the main focus, such as *STAT4* (encoding signal transducer and activator of transcription 4) and the MHC[5,6]. In recent years, the development of genome-wide association study (GWAS) technology has furthered the understanding of disease pathogenesis in hundreds of disorders through the successful identification of new genetic susceptibility loci[7]. With strong evidence of genetic predisposition for primary Sjögren's syndrome[5,8], the application of GWAS is expected to provide new insight into primary Sjögren's syndrome pathogenesis through the identification of new risk loci, the validation of previously explored loci and the discovery of loci shared by primary Sjögren's syndrome and other autoimmune diseases.

Here we carried out a three-stage GWAS of primary Sjögren's syndrome in Han Chinese. In the discovery GWAS stage, we genotyped 642,832 SNPs in 597 primary Sjögren's syndrome cases and 1,090 healthy controls using the Affymetrix Axiom Genome-Wide CHB 1 Array Plate. After quality control filtering for individuals and SNPs (Online Methods), there were 542 cases, 1,050 controls (**Supplementary Table 1**) and 556,134 autosomal SNPs remaining for statistical analysis. Population stratification analysis by principal-component analysis (PCA) showed that all individuals in our study were East Asian and that the cases and controls of Han Chinese ancestry were well matched (**Supplementary Fig. 1**). We conducted SNP-trait association analysis using an additive model in logistic regression with adjustment for sex and age and with PCA-based correction for population stratification. A Manhattan plot of genome-wide *P* values

A full list of affiliations appears at the end of the paper.

Received 19 June; accepted 6 September; published online 6 October 2013; doi:10.1038/ng.2779

Table 1 Summary of association results for three non-MHC SNPs and two representative MHC SNPs associated with primary Sjögren's syndrome in Han Chinese

SNP	Chr.	Position[a]	Allele[b]	Gene	Stage	MAF in cases	MAF in controls	OR (95% CI)	*P* value
rs117026326	7	74126034	T/C	*GTF2I*	Discovery	0.2657	0.1514	2.12 (1.76–2.55)	1.76×10^{-15}
					Replication I	0.2902	0.1622	2.02 (1.69–2.42)	1.76×10^{-14}
					Replication II	0.2903	0.1438	2.38 (2.03–2.80)	3.66×10^{-26}
					Combined	0.2830	0.1501	2.20 (1.99–2.43)	1.31×10^{-53}
rs10168266	2	191935804	T/C	*STAT4*	Discovery	0.4510	0.3317	1.63 (1.40–1.89)	2.77×10^{-10}
					Replication I	0.4212	0.3408	1.41 (1.20–1.66)	2.20×10^{-5}
					Replication II	0.4019	0.3356	1.34 (1.17–1.53)	1.71×10^{-5}
					Combined	0.4234	0.3357	1.44 (1.32–1.57)	1.77×10^{-17}
rs5029939	6	138195723	G/C	*TNFAIP3*	Discovery	0.0638	0.0473	1.41 (1.02–1.94)	3.69×10^{-2}
					Replication I	0.0774	0.0417	1.76 (1.27–2.43)	6.54×10^{-4}
					Replication II	0.0739	0.0443	1.78 (1.36–2.34)	3.16×10^{-5}
					Combined	0.0722	0.0445	1.67 (1.40–1.99)	7.75×10^{-9}
rs9271588	6	32590953	C/T	*HLA-DRB1*, *HLA-DQA1*	Discovery	0.3247	0.4571	0.58 (0.50–0.68)	9.50×10^{-12}
					Replication I	0.3262	0.4739	0.54 (0.46–0.63)	2.92×10^{-14}
					Replication II	0.3468	0.4715	0.60 (0.52–0.69)	3.52×10^{-13}
					Combined	0.3329	0.4681	0.57 (0.53–0.63)	8.52×10^{-37}
rs4282438	6	33072172	G/T	*HLA-DPB1*, *COL11A2*	Discovery	0.4502	0.3445	1.61 (1.37–1.89)	5.97×10^{-9}
					Replication I	0.4632	0.3583	1.65 (1.40–1.94)	1.47×10^{-9}
					Replication II	0.4553	0.3587	1.51 (1.31–1.73)	4.62×10^{-9}
					Combined	0.4566	0.3546	1.58 (1.45–1.72)	8.77×10^{-25}

Chr., chromosome.
[a]Positions are based on human genome version 19 (hg19). [b]Minor/major alleles.

of association is shown in **Supplementary Figure 2**. A quantile-quantile plot of these values showed a deviation only at the far tail of the distribution of observed *P* values from those expected by chance (**Supplementary Fig. 3**). The genomic inflation factor (λ) calculated with or without SNPs in the MHC region was 1.02 and 1.01 before and after PCA-based correction, respectively, indicating that the impact of population stratification was negligible in our study samples. Our genome-wide association analysis identified 18 SNPs with genome-wide significance ($P < 5 \times 10^{-8}$), with *P* values ranging from 1.76×10^{-15} to 3.63×10^{-8}.

The 18 significantly associated SNPs from our discovery stage mapped to 3 loci. One of the SNPs was in an intron of the *GTF2I* gene at 7q11.23 (rs117026326: $P = 1.76 \times 10^{-15}$, OR = 2.12). Two other SNPs, which were in linkage disequilibrium (LD) with each other ($r^2 > 0.8$), were in intronic regions of the gene *STAT4* at 2q32.2-q32.3 (rs10168266: $P = 2.77 \times 10^{-10}$, OR = 1.63; rs7574865: $P = 8.76 \times 10^{-10}$, OR = 1.60). The remaining 15 SNPs were located in the MHC region, and conditional logistic regression analysis (Online Methods) showed that there were 2 independent association signals at 6p21.3 represented by the 2 most significant SNPs, rs9271588 ($P = 9.50 \times 10^{-12}$, OR = 0.58) and rs4282438 ($P = 5.97 \times 10^{-9}$, OR = 1.61). We then chose the representative SNPs with the strongest association in these loci for replication. To discover more susceptibility loci, we selected for replication additional representative non-MHC SNPs with association $P < 5 \times 10^{-5}$ and non-MHC SNPs with association $P < 5 \times 10^{-2}$ but reported to be associated with a Sjögren's syndrome–related autoimmune disease by GWAS (Online Methods). We also analyzed X-chromosomal SNPs, although no X-chromosomal SNP matched the above SNP selection criteria for replication. In total, we included 33 SNPs with clear cluster plots in the follow-up replication I stage (665 cases and 864 controls), which was performed with the iPLEX MassARRAY platform (Sequenom). In replication I, 18 of the 33 SNPs passed platform concordance validation and quality control and showed the same direction of association in replication I as in the discovery stage (Online Methods); these 18 SNPs were taken forward to the replication II stage (638 cases and 1,863 controls) and were genotyped using the iPLEX MassARRAY platform. Together, replications I and II comprised 1,303 cases and 2,727 controls of Han Chinese ancestry (**Supplementary Table 1**).

After quality control for SNPs in replication II, 17 SNPs remained for association analysis. The combined analysis of the 3 stages for these 17 SNPs identified 3 non-MHC loci and 2 loci in the MHC region with genome-wide significant association signals ($P_{combined} < 5 \times 10^{-8}$): *GTF2I* at 7q11.23 (rs117026326: $P_{combined} = 1.31 \times 10^{-53}$, $OR_{combined} = 2.20$, 95% confidence interval (CI) = 1.99–2.43), *STAT4* at 2q32.2-q32.3 (rs10168266: $P_{combined} = 1.77 \times 10^{-17}$, $OR_{combined} = 1.44$, 95% CI = 1.32–1.57), *TNFAIP3* at 6q23 (rs5029939: $P_{combined} = 7.75 \times 10^{-9}$, $OR_{combined} = 1.67$, 95% CI = 1.40–1.99) and the MHC region (rs9271588: $P_{combined} = 8.52 \times 10^{-37}$, $OR_{combined} = 0.57$, 95% CI = 0.53–0.63; rs4282438: $P_{combined} = 8.77 \times 10^{-25}$, $OR_{combined} = 1.58$, 95% CI = 1.45–1.72) (**Table 1**). Results from meta-analysis (Online Methods and **Supplementary Table 2**) showed that there was no stage-related heterogeneity for the five genome-wide significant SNPs (three non-MHC and two MHC) ($P_{Cochran's\ Q} > 0.05$). Variations in genotype frequency between cases and controls for the five SNPs are shown in **Supplementary Figure 4**.

To evaluate whether the index SNPs could completely explain the association signals for their corresponding loci, we performed imputation and SNP-trait association tests by conditional logistic regression (Online Methods). Regional association plots for the three non-MHC susceptibility loci showed that the most significant signal occurred for the index SNPs rs117026326 in *GTF2I* and rs10168266

in *STAT4* but not for rs5029939 in *TNFAIP3* (**Supplementary Fig. 5**). There was no residual association (association after the signal for the index SNP was removed) at these three index SNPs in the three regions. Our epistasis analysis (Online Methods) did not find any significant pairwise interaction between any of the three genome-wide significant SNPs or between the three significant SNPs and other SNPs in the genome. The regional plot for the MHC region showed that the genome-wide significant SNPs (both genotyped and imputed) spanned a ~460-kb region (32.21–32.67 Mb) and an ~80-kb region (33.02–33.10 Mb) within the human leukocyte antigen (HLA) class II region at 6p21.3 (**Supplementary Fig. 6**). The first region contains *C6orf10*, *BTNL2*, *HLA-DRA*, *HLA-DRB5*, *HLA-DRB6*, *HLA-DRB1*, *HLA-DQA1* and *HLA-DQB1*, and the second region contains *HLA-DPA1* and *HLA-DPB1*. The HLA class II region has previously been reported to be associated with primary Sjögren's syndrome[9,10].

For the *GTF2I* locus, the regional plot showed that only one SNP, rs117026326, supported the association signal of this locus, and the corresponding region was sparsely covered by the genome-wide SNP chip. To find additional SNPs to validate the association, we performed fine mapping of an approximately 560-kb genomic region (73.95–74.51 Mb on chromosome 7) using 600 cases and 1,100 controls with the iPLEX MassARRAY platform (Online Methods). The 560-kb region was centered on *GTF2I* and its adjacent gene *NCF1*, which is reported to be involved in inflammation[11]. We selected 41 tag SNPs with minor allele frequency (MAF) >0.05 on the basis of Han Chinese from Beijing, China (CHB) data from the 1000 Genomes Project Integrated Phase 1 Release and successfully genotyped 33 tag SNPs (including 4 SNPs already genotyped in the discovery stage). The 33 tag SNPs were mainly located in the region containing *GTF2IRD1* and *GTF2I*. After imputation (Online Methods), we identified nine SNPs associated with genome-wide significance ($P < 5 \times 10^{-8}$) (**Supplementary Table 3**). Of these nine SNPs, three were genotyped (rs117026326, $P = 3.66 \times 10^{-16}$; rs73366469, $P = 1.74 \times 10^{-14}$; rs80346167, $P = 1.79 \times 10^{-11}$), and six were imputed. SNP rs117026326 showed the strongest association, and there was no residual association detected at it in conditional logistic regression analysis. Seven of the other eight SNPs were in LD ($r^2 > 0.3$, between 0.33 and 0.64) with rs117026326, and one had $r^2 = 0.24$ with rs117026326. The nine SNPs are located in intronic regions of *GTF2IRD1* and *GTF2I* and in the intergenic region between these genes. In addition, one genotyped SNP in the intergenic region between *GTF2IRD1* and *GTF2I* showed association with high significance (rs4717901: $P = 6.36 \times 10^{-8}$, $r^2 = 0.34$ with rs117026326). The fine-mapping results further supported the newly discovered signal at *GTF2I* by confirming statistical evidence of association over multiple SNPs. These findings also extended the region of association from *GTF2I* to *GTF2IRD1-GTF2I*. For *NCF1*, there were a total of two common variants (MAF > 0.05) in this gene identified on the basis of CHB data from the 1000 Genomes Project Integrated Phase 1 Release, and both were selected as tag SNPs for fine mapping. Only one of these two SNPs was successfully genotyped, and no association signal was found at this SNP. The regional association plot for fine mapping is shown in **Figure 1**.

Our study identifies a new susceptibility locus, *GTF2IRD1-GTF2I* (rs117026326), for primary Sjögren's syndrome at 7q11.23. To investigate how common variants at this locus might influence susceptibility, we further explored the biological functions of *GTF2I* and *GTF2IRD1*. Our expression quantitative trait locus (eQTL) analysis (Online Methods) did not find any correlation between genotype at rs117026326 and *GTF2I* (or *GTF2IRD1*) expression. *GTF2I* encodes general transcription factor IIi (TFII-I), which is a multifunctional phosphoprotein with roles in transcription and signal transduction.

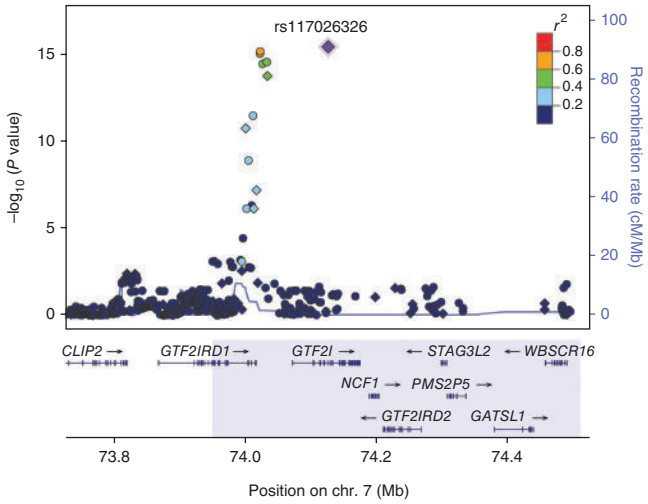

Figure 1 Regional plot of association results for the newly identified primary Sjögren's syndrome susceptibility locus at 7q11.23. Data plotted within the region at 73.95–74.51 Mb (highlighted in transparent blue in the bottom panel) show the results of fine mapping, and data plotted outside the highlighted region show the results of the discovery stage. The association results of both genotyped (diamonds) and imputed (circles) SNPs are shown with the recombination rates (NCBI Build GRCh37, estimated using all populations of HapMap Phase 2: CEU (Utah residents of Northern and Western European ancestry), JPT (Japanese in Tokyo, Japan), CHB (Han Chinese in Beijing, China) and YRI (Yoruba in Ibadan, Nigeria)). The $-\log_{10}$ (*P* values) of SNPs (left *y* axis) are presented according to their chromosomal positions (*x* axis). Genetic recombination rates are represented by light-blue lines, and genes within the regions are shown in the bottom panel. The index SNP rs117026326 (genome-wide significant SNP in combined analysis) is shown as a larger purple diamond labeled by rs number, and its LD (r^2 based on CHB data from the 1000 Genomes Project Integrated Phased 1 Release) with the remaining SNPs is indicated by different colors. The plot was drawn using LocusZoom.

GTF2I has been reported to be one of the main genes responsible for neurocognitive defects in Williams-Beuren syndrome[12]. There have been a number of studies indicating important roles of TFII-I in signal-induced transcriptional regulation in response to various signaling pathways, including immune signaling of both B cells and T cells[13]. The molecular function of *GTF2IRD1* has not been fully studied. It has been reported that *GTF2IRD1* is involved in mammalian craniofacial and cognitive development[14]. Our study also confirmed the primary Sjögren's syndrome susceptibility loci *STAT4* (rs10168266) and *TNFAIP3* (rs5029939). Both *STAT4* and *TNFAIP3* have been reported by candidate gene studies to be associated with primary Sjögren's syndrome in Europeans[5,6,15–18] and have been found by GWAS or candidate gene studies to be associated with several other autoimmune diseases[19–33] (**Supplementary Table 4**).

In summary, our study identified a new locus, *GTF2IRD1-GTF2I*, associated with primary Sjögren's syndrome and confirmed associations of variants in the *STAT4*, *TNFAIP3* and MHC regions in Han Chinese. To our knowledge, this is the first GWAS for primary Sjögren's syndrome. Further resequencing of the *GTF2IRD1-GTF2I* risk haplotype and association mapping of newly identified variants, as well as functional studies of *GTF2IRD1* and *GTF2I*, are expected to help clarify how they are involved in autoimmunity and primary Sjögren's syndrome. As with other autoimmune diseases, primary Sjögren's syndrome appears to be influenced by both genetic susceptibility and environmental triggers. Further studies need to be

performed to address the interaction between genetic risk variants and environmental risk factors, such as infection with Epstein-Barr virus[34]. To provide more information, we have summarized the association results of SNPs with *P* values between 1×10^{-5} and 5×10^{-8} in the discovery GWAS stage (**Supplementary Table 5**) and annotated the candidate loci tagged by these SNPs. Among the candidate loci, *STAT4*, *HCG22*, *C6orf15*, *NOTCH4*, *C6orf10*, *BTNL2*, *HLA-DRA*, *HLA-DQA1*, *HLA-DQB1*, *HLA-DQA2*, *HLA-DPA1*, *HLA-DPB1*, *PRDM1*, *ATG5*, *C7orf72* and *IKZF1* are also reported to be associated with SLE or other autoimmune diseases and deserve further investigation. To enable researchers to have a better understanding of whether our GWAS results provide support for association with previously published candidate genes for Sjögren's syndrome, we summarized published candidate gene studies for Sjögren's syndrome and report our GWAS results for each gene (**Supplementary Table 6**). Because of the close clinical relationship and shared pathophysiology of SLE and primary Sjögren's syndrome[35], we also summarized known SLE risk loci and report our GWAS results for each gene (**Supplementary Table 7**). Our study will help advance the understanding of the etiology of primary Sjögren's syndrome.

URLs. R, http://www.r-project.org/; PLINK, http://pngu.mgh.harvard.edu/~purcell/plink/; EIGENSOFT, http://genetics.med.harvard.edu/reich/Reich_Lab/Software.html; Haploview, http://www.broadinstitute.org/scientific-community/science/programs/medical-and-population-genetics/haploview/haploview; MACH, http://www.sph.umich.edu/csg/abecasis/MACH/index.html; LocusZoom, http://csg.sph.umich.edu/locuszoom/; Genevar, http://www.sanger.ac.uk/resources/software/genevar/; International HapMap Project, http://hapmap.ncbi.nlm.nih.gov/; 1000 Genomes Project, http://www.1000genomes.org/; National Human Genome Research Institute GWAS Catalog, http://www.genome.gov/gwastudies/.

METHODS

Methods and any associated references are available in the online version of the paper.

Note: Any Supplementary Information and Source Data files are available in the online version of the paper.

ACKNOWLEDGMENTS

We thank all participants recruited in this study. This work was supported by grants from the Research Special Fund for Public Welfare Industry of Health (201202004 to Fengchun Zhang), the National Science Technology Pillar Program in the 11th Five-Year Plan (2008BAI59B03 to Fengchun Zhang), the National Program on Key Research Project of New Drug Innovation (2012ZX09303006-002 to Fengchun Zhang), the Knowledge Innovation Program of the Chinese Academy of Sciences (KSCX2-EW-J-8 to J.W.), the CAS/SAFEA International Partnership Program for Creative Research Teams (Y2CX131003 to J.W.) and the National Natural Science Foundation of China (81072486 and 81172857 to Yongzhe Li and 81101545 to K.Z.).

AUTHOR CONTRIBUTIONS

Fengchun Zhang and J.W. conceived and designed the overall project. Yongzhe Li and H.C. directed and managed sample collection and diagnosis. K.Z. supervised data analysis. Yongzhe Li, K.Z., H.C., F.S., Juanjuan Xu and Z.W. managed clinical information and genotyping. K.Z., L. Zhang, Y.D. and J.O. performed statistical data analysis. K.Z. and J.W. wrote the manuscript. J.O., L. Zhang, Yongzhe Li, H.C. and Fengchun Zhang revised the manuscript. The following authors contributed to sample collection: F.S., Juanjuan Xu, Z.W., P.L., H. Luan, Xi Li, L. Wu, H. Li, H.W., Xiangpei Li, Xiaomei Li, Xiao Zhang, L.G., L.D., L.S., X. Zuo, Jianhua Xu, H.G., Z.L., S.T., M.W., Xiaofeng Li, W.X., G.W., P.Z., M.S., S.L., D.Z., W.L., Yi Wang, C.H., Q.J., G.L., B.L., S.H., W. Zhang, Z.Z., X.Y., M.L., W.H., C.Z., X. Leng, L.B., Yongfu Wang, Fengxiao Zhang, Q.S., W.Q., Xuewu Zhang, Y. Jia, J.S., Q.L., Y.H., Q. Wu, D.X., W. Zheng, M.Z., Q. Wang, Y.F., Xuan Zhang, J.L., Y. Jiang, X.T., L. Zhao, L. Wang, B.Z., Yang Li, Y.Z. and X. Zeng. Fengchun Zhang, J.W., Yongzhe Li and K.Z. obtained funding for this study. All authors reviewed the manuscript.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

1. Fox, R.I. Sjogren's syndrome. *Lancet* **366**, 321–331 (2005).
2. Helmick, C.G. *et al.* Estimates of the prevalence of arthritis and other rheumatic conditions in the United States. Part I. *Arthritis Rheum.* **58**, 15–25 (2008).
3. Zhang, N.Z. *et al.* Prevalence of primary Sjogren's syndrome in China. *J. Rheumatol.* **22**, 659–661 (1995).
4. Ice, J.A. *et al.* Genetics of Sjogren's syndrome in the genome-wide association era. *J. Autoimmun.* **39**, 57–63 (2012).
5. Cobb, B.L., Lessard, C.J., Harley, J.B. & Moser, K.L. Genes and Sjogren's syndrome. *Rheum. Dis. Clin. North Am.* **34**, 847–868 (2008).
6. Palomino-Morales, R.J., Diaz-Gallo, L.M., Witte, T., Anaya, J.M. & Martin, J. Influence of *STAT4* polymorphism in primary Sjogren's syndrome. *J. Rheumatol.* **37**, 1016–1019 (2010).
7. Visscher, P.M., Brown, M.A., McCarthy, M.I. & Yang, J. Five years of GWAS discovery. *Am. J. Hum. Genet.* **90**, 7–24 (2012).
8. Williams, P.H. *et al.* Horizons in Sjogren's syndrome genetics. *Clin. Rev. Allergy Immunol.* **32**, 201–209 (2007).
9. Kang, H.I. *et al.* Comparison of HLA class II genes in Caucasoid, Chinese, and Japanese patients with primary Sjogren's syndrome. *J. Immunol.* **150**, 3615–3623 (1993).
10. Gottenberg, J.E. *et al.* In primary Sjogren's syndrome, HLA class II is associated exclusively with autoantibody production and spreading of the autoimmune response. *Arthritis Rheum.* **48**, 2240–2245 (2003).
11. Gill, H.K. *et al.* Defining p47-phox deficient chronic granulomatous disease in a Malay family. *Asian Pac. J. Allergy Immunol.* **30**, 313–320 (2012).
12. Vandeweyer, G., Van der Aa, N., Reyniers, E. & Kooy, R.F. The contribution of *CLIP2* haploinsufficiency to the clinical manifestations of the Williams-Beuren syndrome. *Am. J. Hum. Genet.* **90**, 1071–1078 (2012).
13. Roy, A.L. Biochemistry and biology of the inducible multifunctional transcription factor TFII-I: 10 years later. *Gene* **492**, 32–41 (2012).
14. Tassabehji, M. *et al. GTF2IRD1* in craniofacial development of humans and mice. *Science* **310**, 1184–1187 (2005).
15. Korman, B.D. *et al.* Variant form of *STAT4* is associated with primary Sjogren's syndrome. *Genes Immun.* **9**, 267–270 (2008).
16. Nordmark, G. *et al.* Additive effects of the major risk alleles of *IRF5* and *STAT4* in primary Sjogren's syndrome. *Genes Immun.* **10**, 68–76 (2009).
17. Gestermann, N. *et al. STAT4* is a confirmed genetic risk factor for Sjogren's syndrome and could be involved in type 1 interferon pathway signaling. *Genes Immun.* **11**, 432–438 (2010).
18. Musone, S.L. *et al.* Sequencing of *TNFAIP3* and association of variants with multiple autoimmune diseases. *Genes Immun.* **12**, 176–182 (2011).
19. Han, J.W. *et al.* Genome-wide association study in a Chinese Han population identifies nine new susceptibility loci for systemic lupus erythematosus. *Nat. Genet.* **41**, 1234–1237 (2009).
20. Yang, W. *et al.* Genome-wide association study in Asian populations identifies variants in *ETS1* and *WDFY4* associated with systemic lupus erythematosus. *PLoS Genet.* **6**, e1000841 (2010).
21. Yang, W. *et al.* Meta-analysis followed by replication identifies loci in or near *CDKN1B*, *TET3*, *CD80*, *DRAM1* and *ARID5B* as associated with systemic lupus erythematosus in Asians. *Am. J. Hum. Genet.* **92**, 41–51 (2013).
22. Okada, Y. *et al.* A genome-wide association study identified *AFF1* as a susceptibility locus for systemic lupus eyrthematosus in Japanese. *PLoS Genet.* **8**, e1002455 (2012).
23. Hom, G. *et al.* Association of systemic lupus erythematosus with *C8orf13-BLK* and *ITGAM-ITGAX*. *N. Engl. J. Med.* **358**, 900–909 (2008).
24. Graham, R.R. *et al.* Genetic variants near *TNFAIP3* on 6q23 are associated with systemic lupus erythematosus. *Nat. Genet.* **40**, 1059–1061 (2008).
25. Chung, S.A. *et al.* Differential genetic associations for systemic lupus erythematosus based on anti-dsDNA autoantibody production. *PLoS Genet.* **7**, e1001323 (2011).
26. Zhernakova, A. *et al.* Meta-analysis of genome-wide association studies in celiac disease and rheumatoid arthritis identifies fourteen non-HLA shared loci. *PLoS Genet.* **7**, e1002004 (2011).
27. Radstake, T.R. *et al.* Genome-wide association study of systemic sclerosis identifies *CD247* as a new susceptibility locus. *Nat. Genet.* **42**, 426–429 (2010).
28. Mells, G.F. *et al.* Genome-wide association study identifies 12 new susceptibility loci for primary biliary cirrhosis. *Nat. Genet.* **43**, 329–332 (2011).
29. Shimane, K. *et al.* The association of a nonsynonymous single-nucleotide polymorphism in *TNFAIP3* with systemic lupus erythematosus and rheumatoid arthritis in the Japanese population. *Arthritis Rheum.* **62**, 574–579 (2010).
30. Musone, S.L. *et al.* Multiple polymorphisms in the *TNFAIP3* region are independently associated with systemic lupus erythematosus. *Nat. Genet.* **40**, 1062–1064 (2008).
31. Plenge, R.M. *et al.* Two independent alleles at 6q23 associated with risk of rheumatoid arthritis. *Nat. Genet.* **39**, 1477–1482 (2007).

32. Coenen, M.J. *et al.* Common and different genetic background for rheumatoid arthritis and coeliac disease. *Hum. Mol. Genet.* **18**, 4195–4203 (2009).
33. Trynka, G. *et al.* Coeliac disease–associated risk variants in *TNFAIP3* and *REL* implicate altered NF-κB signalling. *Gut* **58**, 1078–1083 (2009).
34. Toussirot, E. & Roudier, J. Epstein-Barr virus in autoimmune diseases. *Best Pract. Res. Clin. Rheumatol.* **22**, 883–896 (2008).
35. Perl, A. Emerging new pathways of pathogenesis and targets for treatment in systemic lupus erythematosus and Sjogren's syndrome. *Curr. Opin. Rheumatol.* **21**, 443–447 (2009).

[1]Department of Rheumatology and Clinical Immunology, Peking Union Medical College Hospital, Peking Union Medical College and Chinese Academy of Medical Sciences, Key Laboratory of Rheumatology and Clinical Immunology, Ministry of Education, Beijing, China. [2]Key Laboratory of Mental Health, Institute of Psychology, Chinese Academy of Sciences, Beijing, China. [3]Department of Rheumatology and Immunology, The People's Hospital of Xinjiang Uygur Autonomous Region, Urumqi, China. [4]Department of Rheumatology, Affiliated Hospital of Inner Mongolia Medical College, Hohhot, China. [5]Department of Rheumatology, The Second Affiliated Hospital, College of Medicine, Zhejiang University, Hangzhou, China. [6]Department of Rheumatology and Immunity, Affiliated Anhui Provincial Hospital, Anhui Medical University, Hefei, China. [7]Department of Rheumatology and Immunology, Guangdong General Hospital, Guangdong Academy of Medical Sciences, Guangzhou, China. [8]Department of Immunology, General Hospital of Tianjin Medical University, Tianjin, China. [9]Department of Rheumatology, SUN Yat-sen Memorial Hospital of SUN Yat-sen University, Guangzhou, China. [10]Department of Rheumatology and Immunology, The Affiliated Drum Tower Hospital of Nanjing University Medical School, Nanjing, China. [11]Department of Rheumatology and Immunology, Xiangya Hospital Central-South University, Changsha, China. [12]Department of Rheumatology, The First Affiliated Hospital of Anhui Medical University, Hefei, China. [13]Department of Rheumatology and Immunology, Affiliated Heping Hospital, Changzhi Medical College, Changzhi, China. [14]Department of Rheumatology and Immunology, The First Affiliated Hospital of Bengbu Medical College, Bengbu, China. [15]Department of Rheumatology and Immunology, Tangshan Gongren Hospital, Tangshan, China. [16]Department of Rheumatology and Immunology, The First People's Hospital of Changzhou, Changzhou, China. [17]Department of Rheumatology and Immunology, Shanxi Medical University Second Hospital, Taiyuan, China. [18]Department of Rheumatology, The First Hospital, China Medical University, Shenyang, China. [19]Department of Rheumatology and Immunology, China-Japan Friendship Hospital, Beijing, China. [20]Department of Rheumatology and Immunology, The Fourth Military Medical University Xijing Hospital, Xi'an, China. [21]Department of Rheumatology and Immunology, The First Affiliated Hospital of Zhengzhou University, Zhengzhou, China. [22]Department of Rheumatology and Immunology, Changhai Hospital, Second Military Medical University, Shanghai, China. [23]Department of Rheumatology and Immunology, The First Teaching Hospital of Tianjin University of Traditional Chinese Medicine, Tianjin, China. [24]Department of Rheumatology and Immunology, The Second Hospital of Lanzhou University, Lanzhou, China. [25]Department of Rheumatology and Immunology, Beijing Hospital, Beijing, China. [26]Department of Rheumatology, Guang'anmen Hospital, China Academy of Chinese Medical Sciences, Beijing, China. [27]Department of Laboratory, Guang'anmen Hospital, China Academy of Chinese Medical Sciences, Beijing, China. [28]Department of Rheumatology and Immunology, The Affiliated Hospital of Medical College Qingdao University, Qingdao, China. [29]Department of Immunology and Rheumatology, Tongji Hospital, Tongji Medical College, Huazhong University of Science & Technology, Wuhan, China. [30]Department of Rheumatology and Immunology, Peking University First Hospital, Beijing, China. [31]Department of Traditional Chinese Medicine, Peking Union Medical College Hospital, Peking Union Medical College and Chinese Academy of Medical Sciences, Beijing, China. [32]Department of Rheumatology and Immunology, Guangxi Medical University First Affiliated Hospital, Nanning, China. [33]Department of Rheumatology and Immunology, China-Japan Union Hospital of Jilin University, Changchun, China. [34]Department of Rheumatology and Immunology, The First Affiliated Hospital of Baotou Medical College, Baotou, China. [35]Department of Rheumatology and Immunology, Hebei People's Hospital, Shijiazhuang, China. [36]Department of Rheumatology and Immunology, Tianjin First Central Hospital, Tianjin, China. [37]Department of Rheumatology and Immunology, Peking University People's Hospital, Beijing, China. [38]Department of Rheumatology and Immunology, First People's Hospital of Yunnan University, Kunming, China. [39]Department of Rheumatology, Jiangsu Provincial People's Hospital, Nanjing, China. [40]Department of Rheumatology and Immunology, Sichuan Academy of Medical Sciences & Sichuan Provincial People's Hospital, Chengdu, China. [41]Department of Rheumatology and Immunology, The Second Affiliated Hospital of Harbin Medical University, Harbin, China. [42]These authors contributed equally to this work. [43]These authors jointly directed this work. Correspondence should be addressed to Fengchun Zhang (zhangfccra@aliyun.com) or J.W. (wangjing@psych.ac.cn).

ONLINE METHODS

Study samples. We used a three-stage study design. The discovery GWAS stage included 597 primary Sjögren's syndrome cases and 1,090 healthy controls of Han Chinese origin. The replication I and II stages included 1,303 cases and 2,727 controls of Han Chinese origin. In fine mapping, we used 600 cases and 1,100 controls (all Han Chinese), of whom 496 cases and 530 controls were discovery GWAS case-control samples and 104 cases and 570 controls were newly collected because some of the GWAS DNA samples were not available (**Supplementary Table 1**). All cases were recruited through the cooperation of 40 centers in China (the list of centers is included in authors' affiliations). All cases were diagnosed with primary Sjögren's syndrome by at least two rheumatologists according to the American-European Consensus Group (AECG) criteria for primary Sjögren's syndrome[36] and were not diagnosed with any other autoimmune disease. All controls were recruited from the Peking Union Medical College Hospital according to the following rules: (i) no significant history of disease; (ii) no family history of rheumatologic diseases; (iii) normal biochemical and immunological profile; and (iv) negative serology for anti-Ro (SSA) and anti-La (SSB) antibodies. Informed consent was given by each individual at recruitment. Approval for our study was obtained from the institutional review board of Peking Union Medical College Hospital. For all case and control individuals recruited in our study, anti-Ro (SSA) and anti-La (SSB) antibodies were detected by ELISA (EUROIMMUN), and genomic DNA was extracted from peripheral blood samples using a whole-blood DNA extraction kit (solution type) (BioTeke).

Genotyping and quality control in GWAS. Genome-wide genotyping analysis was conducted using the Affymetrix Axiom Genome-Wide CHB 1 Array Plate with 642,832 SNPs at CapitalBio Co., Ltd. Genome-wide genotypes were called by the Axiom GT1 algorithm. Systemic quality control was then performed for individuals and SNPs. Stepwise quality control for the 597 cases and 1,090 controls at the discovery stage was to remove individuals (i) with Affymetrix Dish quality control (DQC) <0.82 (following Affymetrix's data analysis guideline) (n = 0); (ii) with per-individual call rate <95% (n = 19); (iii) with per-individual autosomal heterozygosity >5 s.d. away from the mean[37] (n = 14); (iv) with the wrong assigned sex (n = 11); (v) who were one of a pair of individuals with a cryptic relationship closer than a third-degree relative (proportion identity by descent (IBD) PI_HAT ≥ 0.125) or unexpected duplication, i.e., identity >98% (the one with the lower call rate was removed in this case) (n = 39); and (vi) who did not have age information (n = 12). After applying these quality control steps, 542 cases and 1,050 controls remained. The remaining samples were assessed for population stratification using PCA implemented in EIGENSOFT4.2 (refs. 38,39) with 209 HapMap subjects as the reference panel (60 CEU, 60 YRI, 45 CHB and 44 JPT)[40]. A total of 22,112 genotyped autosomal SNPs, which were in low LD (MAF > 0.35 and r^2 < 0.05 for each pair of SNPs)[41,42] and absent from the 5 long-range LD regions[43], were included in the PCA. The format of EIGENSOFT output (eigenvectors, i.e., principal components) was adapted to R 2.8.1 (see URLs) for plotting, and the first two eigenvectors were plotted. Results confirmed that our samples were of East Asian ancestry and that no population outlier (>6 s.d. from the mean for any of the top ten eigenvectors) was present. Population stratification was also assessed by only using the case-control data, and the Tracy-Widom test was employed to detect significant eigenvectors (P < 0.05). We found that the first eigenvector (eigenvector 1) was significant, and it was used as one of the covariates in the subsequent statistical analysis. A summary of the individuals who passed quality control is shown in **Supplementary Table 1**. After quality control for individuals, quality control for SNPs was performed to remove SNPs (i) with per-SNP call rate <98% either in cases or controls; (ii) with Hardy-Weinberg equilibrium test P < 0.001 for controls; (iii) with MAF <1% either in cases or controls; (iv) with significantly different call rates between cases and controls ($P < 1 \times 10^{-5}$, PLINK command –test-missing)[44]; and (v) not mapped to autosomes. In total, there were 556,134 SNPs remaining after quality control.

SNP selection for replication. For SNPs in the non-MHC regions with $P < 5 \times 10^{-5}$, we grouped SNPs on the basis of their LD information and physical location. Any two SNPs with r^2 > 0.3 (LD was calculated within a distance of 1 Mb), together with the SNPs located between them, were placed into the same group, so that each group represented an approximately independent region. For each region, the SNP with the lowest P value was selected for replication (n = 23). We also selected the two representative SNPs (rs9271588 and rs4282438) for the MHC region (n = 2). Besides these SNPs, additional non-MHC SNPs that showed an association signal with primary Sjögren's syndrome ($P < 5 \times 10^{-2}$) and that had been identified by GWAS (according to the NHGRI GWAS Catalog; see URLs)[45] to be associated with rheumatoid arthritis, SLE, systemic sclerosis or primary biliary cirrhosis (data accessed by 30 September 2012) were selected for replication (n = 8). In total, 33 SNPs were selected for the replication I stage, and all these SNPs showed clear cluster plots (analyzed by Affymetrix Genotyping Console). We selected 18 of the 33 SNPs for a replication II stage.

Genotyping and quality control in replication studies. Genotyping at replication stages was performed using the iPLEX MassARRAY platform (Sequenom). To validate the concordance between the Affymetrix and Sequenom platforms, we randomly selected 120 individuals who passed quality control (60 cases and 60 controls) genotyped during the discovery stage and regenotyped the 33 SNPs selected for replication I using MassARRAY. SNPs with per-SNP concordance < 98% were excluded from further analyses. For quality control during replication stages I and II, we excluded SNPs with (i) call rate < 95% either in cases or controls; (ii) Hardy-Weinberg equilibrium test P < 0.001 in controls; (iii) MAF < 1% either in cases or controls; and (iv) significantly different call rates between cases and controls ($P < 1 \times 10^{-5}$, PLINK command –test-missing). After platform concordance validation and quality control in replication I, 18 of 33 SNPs, which had the same directions of association in the discovery GWAS and replication I stage, were selected for replication II. After quality control on genotype data in replication II, 17 SNPs remained for a combined analysis.

Statistical analysis. Association testing of SNPs with primary Sjögren's syndrome risk in the GWAS, replication stages and combined analysis was carried out by PLINK v1.07 (ref. 46) using an additive model in logistic regression. During the discovery stage, SNP-trait association P values were calculated with sex, age and eigenvector 1 of PCA as covariates. The Manhattan plot of genome-wide $-\log_{10} P$ was generated with Haploview v4.1 (ref. 47), and the quantile-quantile plot of genome-wide $-\log_{10} P$ was generated with R 2.8.1 (see URLs). The genomic control inflation factor λ was calculated with R. Conditional logistic regression (PLINK command –condition) was employed to detect independent association signals in the MHC region and residual association signals in the three non-MHC loci by calculating SNP P values conditioned upon the most significant SNP. All independent SNPs were then combined in one regression model with adjustment for sex and age and PCA-based correction to calculate ORs and P values. Plots of variation in genotype frequency between cases and controls for the genome-wide significant SNPs were also generated with R. Association analyses for the replication stages were implemented with sex and age as covariates, and, in combined analysis, besides sex and age, eigenvector 1 and an indicator of stage were used as covariates (eigenvector 1 was set to zero for the data from replications I and II). Moreover, meta-analysis was conducted by PLINK for the SNPs with genome-wide significance, with Cochran's Q test P value and heterogeneity index I^2 calculated to test heterogeneity by stage. Epistasis tests (PLINK) and interaction tests performed by partitioning χ^2 values[48] were both carried out to detect SNPs interacting with the three non-MHC genome-wide significant SNPs ($P_{\text{interaction}} < 5 \times 10^{-8}$). In examining pairwise interaction of the three SNPs, all $P_{\text{interaction}}$ values exceeded 5×10^{-2} after Bonferroni correction for the three SNP pairs. All association test P values were two-sided and are reported without correction for multiple testing. Genome-wide significance was considered as $P < 5 \times 10^{-8}$ as generally used in GWAS.

Imputation and regional association plot. We used MACH 1.0 (ref. 49) to impute non-genotyped SNPs using the CHB data (97 individuals) from the 1000 Genomes Project Integrated Phase 1 Release[50] as the reference panel. Only imputed SNPs with squared correlation between imputed and true genotypes (Rsq) >0.5 were kept, and imputation results (dosage data) were analyzed by mach2dat[49]. Regional association results were plotted by LocusZoom (stand-alone version)[51].

doi:10.1038/ng.2779

Fine mapping. To define the exact region for fine mapping, we started from the region of *GTF2I* and its adjacent gene *NCF1*. These two genes and their overlapping LD blocks represent a 220-kb region (74.00–74.22 Mb). To remedy the low coverage of the genome-wide SNP chip in the region, we further extended the 220-kb region to a 560-kb region (73.95–74.51 Mb) that included all LD blocks containing SNPs that had $r^2 > 0.3$ with any of the SNPs in the 220-kb region. Haploview was used to select tag SNPs with MAF > 0.05, using Paul de Bakker's Tagger algorithm with threshold $r^2 > 0.9$ on the basis of CHB data from the 1000 Genomes Project Integrated Phase 1 Release. SNPs with problems in primer design were excluded. Genotyping was performed with the iPLEX MassARRAY platform. Quality control for SNPs was the same as that used in the replication stages. For genome-wide significant SNPs identified in the three-stage GWAS and fine mapping, we examined cluster plots and intensity plots from each stage and investigated MAF across plates. All plots were very clear (**Supplementary Fig. 7**), and the mean MAF values across plates were very close to the MAF values from the overall plates, with very little standard deviation (**Supplementary Table 8**).

eQTL analysis. eQTL analysis was performed using Genevar (GENe Expression VARiation) 3.2.0 (ref. 52). Spearman's rank correlation coefficient (ρ) was measured to explore the association between SNP genotype and gene expression for the genome-wide significant SNPs and their corresponding or nearby (± 1 Mb from the SNP) genes on the basis of gene expression data for lymphoblastoid cell lines from 77 HapMap 3 CHB individuals[53] and the corresponding genotypes of these individuals in 1000 Genomes Project Integrated Phase 1 Release data[50]. *P* values were generated from 10,000 permutation samples, and the *P*-value threshold was set to 0.001.

36. Vitali, C. *et al.* Classification criteria for Sjogren's syndrome: a revised version of the European criteria proposed by the American-European Consensus Group. *Ann. Rheum. Dis.* **61**, 554–558 (2002).
37. Hancock, D.B. *et al.* Meta-analyses of genome-wide association studies identify multiple loci associated with pulmonary function. *Nat. Genet.* **42**, 45–52 (2010).
38. Patterson, N., Price, A.L. & Reich, D. Population structure and eigenanalysis. *PLoS Genet.* **2**, e190 (2006).
39. Price, A.L. *et al.* Principal components analysis corrects for stratification in genome-wide association studies. *Nat. Genet.* **38**, 904–909 (2006).
40. Altshuler, D.M. *et al.* Integrating common and rare genetic variation in diverse human populations. *Nature* **467**, 52–58 (2010).
41. Shi, Y. *et al.* A genome-wide association study identifies new susceptibility loci for non-cardia gastric cancer at 3q13.31 and 5p13.1. *Nat. Genet.* **43**, 1215–1218 (2011).
42. Hu, Z. *et al.* A genome-wide association study identifies two new lung cancer susceptibility loci at 13q12.12 and 22q12.2 in Han Chinese. *Nat. Genet.* **43**, 792–796 (2011).
43. Chen, J. *et al.* Genetic structure of the Han Chinese population revealed by genome-wide SNP variation. *Am. J. Hum. Genet.* **85**, 775–785 (2009).
44. Anderson, C.A. *et al.* Data quality control in genetic case-control association studies. *Nat. Protoc.* **5**, 1564–1573 (2010).
45. Hindorff, L.A. *et al.* Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc. Natl. Acad. Sci. USA* **106**, 9362–9367 (2009).
46. Purcell, S. *et al.* PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* **81**, 559–575 (2007).
47. Barrett, J.C., Fry, B., Maller, J. & Daly, M.J. Haploview: analysis and visualization of LD and haplotype maps. *Bioinformatics* **21**, 263–265 (2005).
48. Yang, Y., He, C. & Ott, J. Testing association with interactions by partitioning chi-squares. *Ann. Hum. Genet.* **73**, 109–117 (2009).
49. Li, Y., Willer, C.J., Ding, J., Scheet, P. & Abecasis, G.R. MaCH: using sequence and genotype data to estimate haplotypes and unobserved genotypes. *Genet. Epidemiol.* **34**, 816–834 (2010).
50. Abecasis, G.R. *et al.* An integrated map of genetic variation from 1,092 human genomes. *Nature* **491**, 56–65 (2012).
51. Pruim, R.J. *et al.* LocusZoom: regional visualization of genome-wide association scan results. *Bioinformatics* **26**, 2336–2337 (2010).
52. Yang, T.P. *et al.* Genevar: a database and Java application for the analysis and visualization of SNP-gene associations in eQTL studies. *Bioinformatics* **26**, 2474–2476 (2010).
53. Stranger, B.E. *et al.* Patterns of *cis* regulatory variation in diverse human populations. *PLoS Genet.* **8**, e1002639 (2012).