Supplementary Information

High-density genotyping of immune-related loci identifies new SLE risk variants in individuals with Asian ancestry

Celi Sun[1,24], Julio E. Molineros[1,24], Loren L. Looger[2,24], Xu-jie Zhou[3,24], Kwangwoo Kim[4,24], Yukinori Okada[5,6], Jianyang Ma[7], Yuan-yuan Qi[3], Xana Kim-Howard[1], Prasenjeet Motghare[1], Krishna Bhattarai[1], Adam Adler[1], So-Young Bang[4], Hye-Soon Lee[4], Tae-Hwan Kim[4], Young Mo Kang[8], Chang-Hee Suh[9], Won Tae Chung[10], Yong-Beom Park[11], Jung-Yoon Choe[12], Seung Cheol Shim[13], Yuta Kochi[14], Akari Suzuki[14], Michiaki Kubo[15], Takayuki Sumida[16], Kazuhiko Yamamoto[14,17], Shin-Seok Lee[18], Young Jin Kim[19], Bok-Ghee Han[19], Mikhail Dozmorov[20], Kenneth M. Kaufman[21], Jonathan D. Wren[1], John B. Harley[21], Nan Shen[7,22], Kek Heng Chua[23], Hong Zhang[3], Sang-Cheol Bae[4], Swapan K. Nath[1]

[1] Arthritis and Clinical Immunology Research Program, Oklahoma Medical Research Foundation, Oklahoma City, OK, USA
[2] Howard Hughes Medical Institute, Janelia Research Campus, Ashburn, VA, USA
[3] Renal Division, Peking University First Hospital, Peking University Institute of Nephrology, Key Laboratory of Renal Disease, Ministry of Health of China, and Key Laboratory of Chronic Kidney Disease Prevention and Treatment (Peking University), Ministry of Education, Beijing, China
[4] Department of Rheumatology, Hanyang University Hospital for Rheumatic Diseases, Seoul, Korea
[5] Department of Human Genetics and Disease Diversity, Graduate School of Medical and Dental Sciences, Tokyo Medical and Dental University, Tokyo 113-8510, Japan
[6] Laboratory for Statistical Analysis, RIKEN Center for Integrative Medical Sciences, Yokohama 230-0045, Japan
[7] Joint Molecular Rheumatology Laboratory of the Institute of Health Sciences and Shanghai Renji Hospital, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, and Shanghai Jiaotong University School of Medicine, Shanghai 200025, China
[8] Kyungpook National University Hospital, Daegu, Korea
[9] Ajou University Hospital, Suwon, Korea
[10] Dong-A University Hospital, Department of Internal Medicine, Busan, Korea
[11] Department of Internal Medicine, Yonsei University College of Medicine, Seoul, Korea
[12] Daegu Catholic University Hospital, Daegu, Korea
[13] Daejeon Rheumatoid & Degenerative Arthritis Center, Chungnam National University Hospital, Daejeon, Korea
[14] Laboratory for Autoimmune Diseases, Center for Integrative Medical Sciences, RIKEN, Yokohama 230-0045, Japan
[15] Laboratory for Genotyping Development, RIKEN Center for Integrative Medical Sciences, Yokohama 230-0045, Japan
[16] Department of Internal Medicine, Faculty of Medicine, University of Tsukuba, Tsukuba 305-8575, Japan
[17] Department of Allergy and Rheumatology, Graduate School of Medicine, the University of Tokyo, Tokyo 113-0033, Japan
[18] Chonnam National University Hospital, Gwangju, Korea
[19] Korea National Institute of Health, Osong, Korea
[20] Department of Biostatistics, Virginia Commonwealth University, Richmond, VA, USA
[21] Cincinnati Children's Hospital Medical Center and US Department of Veterans Affairs Medical Center, Cincinnati, OH, USA

[22] Center for Autoimmune Genomics and Etiology (CAGE), Cincinnati Children's Hospital Medical Center, Cincinnati Ohio 45229, USA
[23] Department of Biomedical Science, Faculty of Medicine, University of Malaya, 50603 Kuala Lumpur, Malaysia
[24] These authors contributed equally

Towhom correspondance should be addressed: naths@omrf.org (S.K.N); scbae@hanyang.ac.kr (S.C.B).

Supplementary Note

Study participants

This study included Koreans (KR) recruited from a hospital network run by Hanyang University Hospital for Rheumatic Diseases and 7 other hospitals in South Korea (1843 SLE patients and 3262 controls); Han Chinese (HC) recruited from Peking University in Beijing, China (500 patients and 500 controls); and Malaysian Chinese (MC) recruited from the Health Center of the University of Malaya (302 patients and 296 controls, Malaysians of Chinese ancestry)(**Supplementary Table 1**). All patients satisfied American College of Rheumatology classification for SLE[90,91] based on review of medical records. Controls were recruited from the same geographic regions as SLE cases. Participants provided written consent at study enrollment, and the Internal Review Boards (IRBs) or ethical committees of participating institutions approved this study. Potential identifying information was removed for all participants.

To increase statistical power we incorporated an independent KR out-of-study control GWAS dataset (genotyped using Illumina HumanOmni1-Quad BeadChip GWAS array at the Korea National Institute of Health, which was used in a previous rheumatoid arthritis case-control study[21]). We only used GWAS data (Illumina OMNI-1) from Koreans, since Koreans are incredibly genetically homogenous and constitute our largest cohort. The addition of out-of study controls boosted the statistical power by 12-16% for SNPs with MAF>0.15 and $P<5 \times 10^{-8}$ (**Supplementary Fig. 2**).

The *in silico* replication samples are from a Japanese GWAS dataset (891 cases and 3384 controls)[59]; the other two replication samples are from two independent Han Chinese cohorts (BHC and SHC) recruited at the Peking University First Hospital (601 cases and 1034 controls), and through the Joint Molecular Rheumatology Laboratory of the Institute of Health Sciences and Shanghai Renji Hospital, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, Shanghai Jiaotong University School of Medicine (501 cases and 622 controls), respectively (**Supplementary Table 1**).

Genotyping and quality control

A total of 2645 cases and 4058 controls from three Asian populations were genotyped using the Illumina ImmunoChip array[9] at the Oklahoma Medical Research Foundation (OMRF). Subjects were excluded from analysis if they had >10% missing genotyping. SNPs with overlapping clusters were filtered out before further QC was performed. SNPs were also excluded if they had >5% missing genotyping, were out of Hardy-Weinberg equilibrium (P<0.0001 in controls) or had <0.5% minor allele frequency (MAF). All QC'd SNPs (**Supplementary Table 2**) for ImmunoChip datasets were aligned with NCBI Genome Build 37 and oriented to the forward/+ strand. Excluding known SLE-risk loci, there were 11,234 common, independent (*i.e.*, LD $r^2<0.3$) SNPs across all ImmunoChip and out-of-study control datasets. We used this SNP set for all QC on the samples. We calculated the pair-wise relationship between individuals for each cohort using GCTA[92] (estimating the genetic relationship matrix, grm). Individuals with grm>25% were excluded in subsequent analysis due to possible cryptic relatedness. We identified ancestral outliers as those exceeding six standard deviations beyond the mean for the first three principal components (**Supplementary Fig. 15**) obtained by EIGENSTRAT[93] PCA. After QC, we removed 160 cases (133 KR, 10 HC, 17 MC) and 111 controls (95 KR, 7 HC, 9 MC) due

to low call rate, relatedness and outliers, and integrated 3669 KR into this study. In order to examine genomic inflation due to population stratification, we used the same set of SNPs to estimate genomic inflation (λ) and create Q-Q plots (**Supplementary Fig. 16**). We compared genomic inflation across three cohorts by estimating the sample size-corrected λ_{1000}[94]. There was little evidence of stratification, measured by the genomic control inflation factor (λ_{GC}: λ_{KR}=1.12, λ_{HC}=1.07, λ_{MC}=1.02) and sample size-corrected inflation factor[94] (λ_{1000KR}=1.04; λ_{1000HC}=1.09; λ_{1000MC}=1.07).

Several signals were only genotyped on the ImmunoChip array, for various reasons. *GTF2IRD1-GTF2I* was analyzed using only ImmunoChip because of low SNP density and imputation quality in the out-of-study controls' GWAS array. ImmunoChip data alone was also used for *SYNGR1*, due to low quality imputation.

To confirm the strongest novel signal within the *GTF2IRD1-GTF2I* region, we used TaqMan assays to genotype 2-6 promising SNPs (including rs73366469) on a subset of ImmunoChip-genotyped samples. We genotyped 6 SNPs in 1363 KR (624 cases and 739 controls), 4 SNPs in 322 MC (141 cases and 181 controls) at OMRF, and 2 SNPs in 941 HC (477 cases and 464 controls). We also used the Sequenom array to genotype 4 SNPs in both 1635 BHC (601 cases and 1034 controls), and 1123 SHC (501 cases and 622 controls) at Peking University First Hospital, China (**Supplementary Table 6**).

Replication of novel loci

In order to confirm our novel loci, we used data from three replication cohorts. The *in silico* replication study of the Japanese population was conducted by RIKEN Center for Integrative Medical Sciences, Japan. The GWAS data of 891 SLE cases and 3,384 controls were used. SLE cases were collected under the support of the autoimmune disease study group of Research in Intractable Diseases, Japanese Ministry of Health, Labor and Welfare. The control subjects were collected by BioBank Japan Project. Detailed characteristics of the subjects and the study design are described elsewhere[59].

Japanese SLE cases and controls were genotyped using Illumina HumanHap610-Quad and Illumina HumanHap550v3 Genotyping BeadChips (Illumina, CA, USA), respectively. Details of the QC filters (sample call rates $\geq$ 0.98, SNP call rates $\geq$ 0.99, exclusions of closely related subjects and outliers in terms of PCA analysis, exclusion of SNPs with HWE-P $< 1.0 \times 10^{-6}$, SNPs with MAF < 0.01) are described elsewhere[59]. We applied whole-genome genotype imputation of the GWAS data using MiniMac2 software[95,96]. We adopted the East-Asian subjects of 1000 Genomes Project Phase I (a) data as an imputation reference. We excluded the imputed SNPs with MAF < 0.01 or Rsq values < 0.5 (**Supplementary Table 27**).

For our two other replication sets, we genotyped 22 SNPs (**Supplementary Table 28**) on a custom Sequenom array for two independent Han Chinese cohorts from Beijing (BHC) and Shanghai (SHC).

Analysis of the HLA locus

In order to dissect HLA associations, we imputed SNPs, 2- and 4-digit classical alleles, as well as amino-acid residues for eight HLA genes (*A, B, C, DRB1, DQA1, DQB1, DPA1, DPB1*) in KR, HC and MC using SNP2HLA[97]. We used the expanded Asian reference panel described in Kim et al.[22], which

included genotyping data for SNPs and HLA variants for 854 individuals of Korean, Southern Han Chinese, Singapore Chinese, Southern Malaya and Tamil Indian descent[22].

SNPs, amino-acid residues and 2- and 4-digit alleles with low imputation quality (info <0.70) were excluded from the analysis. SLE association analysis for the imputed markers was performed as previously described[98]. In brief, logistic regression and log-likelihood ratio test ("omnibus" test comparing log-likelihood of the null and alternative logistic regression models) were used to test associations of bi-allelic markers and multi-residue amino-acid positions, respectively. The fitted logistic regression models included the first three principal components and a dummy variable to account for the differences between cohorts. To identify independently associated variants within the extended MHC region in the merged set, we performed step-wise conditional logistic regression. In this approach, we started with the most significant variant and added new variants as covariates, until the significance level of the remaining variants reached $>5 \times 10^{-8}$. For example, when we tested the secondary association of amino-acid positions in *HLA-DRB1*, we conditioned on all residues for each position (excluding the most frequent residue, to avoid collinearity). Similarly, when we investigated the secondary effects outside of particular genes (e.g., *HLA-DRB1*), we conditioned on all 4-digit classical alleles of *DRB1* and examined the remaining variant P-values (new variants would have been included if their $P < 5 \times 10^{-8}$).

References

90.   Tan, E.M. *et al.* The 1982 revised criteria for the classification of systemic lupus erythematosus. *Arthritis Rheum* **25**, 1271-7 (1982).
91.   Hochberg, M.C. Updating the American College of Rheumatology revised criteria for the classification of systemic lupus erythematosus. *Arthritis Rheum* **40**, 1725 (1997).
92.   Yang, J., Lee, S.H., Goddard, M.E. & Visscher, P.M. GCTA: a tool for genome-wide complex trait analysis. *Am J Hum Genet* **88**, 76-82 (2011).
93.   Price, A.L. *et al.* Principal components analysis corrects for stratification in genome-wide association studies. *Nat Genet* **38**, 904-9 (2006).
94.   Freedman, M.L. *et al.* Assessing the impact of population stratification on genetic association studies. *Nat Genet* **36**, 388-93 (2004).
95.   Fuchsberger, C., Abecasis, G.R. & Hinds, D.A. minimac2: faster genotype imputation. *Bioinformatics* **31**, 782-4 (2015).
96.   Howie, B., Fuchsberger, C., Stephens, M., Marchini, J. & Abecasis, G.R. Fast and accurate genotype imputation in genome-wide association studies through pre-phasing. *Nat Genet* **44**, 955-9 (2012).
97.   Jia, X. *et al.* Imputing amino acid polymorphisms in human leukocyte antigens. *PLoS One* **8**, e64683 (2013).
98.   Kim, K., Bang, S.Y., Lee, H.S. & Bae, S.C. Construction and application of a Korean reference panel for imputing classical alleles and amino acids of human leukocyte antigen genes. *PLoS One* **9**, e112546 (2014).

SUPPLEMENTARY MATERIAL

Supplementary Fig 1. Venn diagram for 578 imputation regions with P<5x10^{-3} from three populations' ImmunoChip genotyping. KR=Korean; HC= Han Chinese; MC=Malaya Chinese. Numbers represent the number of regions of the genome with association P <5x10^{-3} prior to imputation.

Supplementary Fig 2. Statistical power for combined cases and controls, before and after addition of out-of-study KR controls. It is of note that our initial statistical power comparison between ImmunoChip-only and ImmunoChip + Korean GWAS data sets showed that for MAF > 0.15, adding controls boosted statistical power by 12-16%. For example, using a cutoff at P ≤ 5x10^{-8} and OR = 1.30, the power increased by 15% (from 0.67 to 0.82) at MAF = 0.30.

Supplementary Fig 3. Regional association plots for 10 novel SLE loci. Plots of -log10(Discovery P-meta) versus position in megabase pairs. The lead SNP is highlighted in purple, while the linkage disequilibrium of the rest of the variants (r2) to the lead SNP is shown in the color legend. a. *TERT*; b. *TCF7*; c. *IL12B*; d. *DEF6*; e. *GTF2I*; f. *PCNXL3* ; g. *RASGRP1*; h. *CD226*; i. *SIGLEC6*; j. *SYNGR1*

Supplementary Fig 4. Epigenetic annotation of novel SLE-associated regions. ENCODE tracks for DNase hypersensitivity, HMM promoter/enhancer/insulator/transcription classification, chromatin marks (CTCF, EZH2, H2Az, H3K27Ac, H3K27me3, H3K6me3, H3K4me1, H3K4me2, H3K4me3, H3K79me2, H3K9Ac, H3K9me3, H4K20me1), transcription factor binding sites and RNA-binding protein sites are shown. Data is taken from the ENCODE tracks for GM12878 cells. For chromatin marks and transcription factor/RNA-binding protein sites, shading of the box indicates intensity of the signal, with black meaning maximum signal. The SNPs tested in this study are shown at the top, with the height of the bar (y-axis on left) indicating -log$_{10}$(P-value). Chromosome position is indicated on the x-axis. For each locus, a blow-up of the region containing the lead SNP and other potentially functional SNPs is shown in a second panel, with key SNPs labeled.

Supplementary Fig 5. Long-range chromatin interactions assessed through Hi-C and ChIA-PET from nine cell lines. Long range interaction strength is shown in the thickness of the connecting lines.

Supplementary Fig 6. Regional association plots for known SLE loci (35 known non-*HLA* genes). Plots of -log10(Discovery P-meta) versus position in megabase pairs. The lead SNP is highlighted in purple, while the linkage of the rest of the variants (r2) to the lead SNP is shown in the color legend. *a STAT4*; b *ETS1*; c *TNFAIP3*; d *TNFSF4*; e *IRF5*; f *IKZF1*; g *BLK*; h *HIP1*; i *IRAK1*; j *WDFY4*; k *IRF8*; l *TNIP1*; m *RASGRP3*; n *PRDM1_ATG5*; o *ARHGAP31_TMEM39A_CD80*; p *SLC15A4*; q *UBE2L3*; r *BANK1*; s *CLEC16A*; t *PDHX_CD44*; u *SNRPC*; v *LBH*; w *DDX6_CXCR5*; x *FCGR2B*; y *PTTG1*; z *FCRL5*; aa *BACH2*; ab *PRKCB*; ac *NCF2*; ad *CSK*; ae *JAZF1*; af *PHRF1*; ag *TLR7*; ah *UBE2E3*; ai *IL4*.

Supplementary Fig 7. Regional association plot for the extended MHC/HLA region. Plots of -log10(Discovery P-meta) versus position in megabase pairs. The lead SNP is highlighted in purple, while the linkage of the rest of the variants (r2) to the lead SNP is shown in the color legend. a) SNP based association analysis; b) Imputed HLA variants (classical allele and residue) association; c) After conditioning for the effect of *HLA-DRB1*, no association remained. We positioned a red dotted line at

the genome-wide significant level (P<5x10-8), as well as a suggestive association line ($5x10^{-8}$<P<$5x10^{-5}$).

Supplementary Fig 8. Enrichment analysis of the novel and SLE loci in cell-specific gene ontology categories. We estimated enrichment of our gene-set in the Gene Ontology database. Overrepresented cell types have a high correlation (Pearson's correlation coefficient) of SLE-loci expression (dark red). P-values (blue bars) that passed the multiple testing threshold (black line) show significant enrichment in SLE loci.

Supplementary Fig 9. Induced network module analysis. In order to investigate how our novel loci fit together, we created an induced network connecting our set of novel loci to interacting genes (blue), proteins (yellow) and drug targets (red).

Supplementary Figure 10. Protein-protein interactions for each novel SLE locus. a *TERT*; b. *TCF7*; c. *IL12B*; d. *DEF6*; e. *GTF2I; f. PCNXL3; g. SIGLEC6*

Supplementary Fig 11. Network illustrating published relationships between known loci (blue) and novel loci (green). Relationships are defined as the co-occurrence of gene names (or their synonyms) within the same abstract. Line thickness is proportional to the number of times the genes were mentioned together. Node size is proportional to the number of connections within this network (i.e., bigger = more connections). Genes that were not co-mentioned with other genes on the list are shown at the left, divided according to how much published information exists for them (poorly $\leq$ 15 papers mentioning the gene in any context, moderately = between 16 and 50 papers, well-annotated $\geq$100 papers).

Supplementary Fig 12. Protein-protein interaction sub-groups identified through DAPPLE analysis of reported biochemical interactions.

Supplementary Fig 13. Enrichment analysis of the novel and known SLE loci in cell type-specific gene expression. We estimated enrichment of our gene-set in human (a: GeneAtlas) and mouse (b: ImmGen) cell lines. Overrepresented cell types have a high correlation (Pearson's correlation coefficient) of SLE-loci expression (dark red). P-values (blue bars) that passed the multiple testing threshold (black line) show significant enrichment in SLE loci.

Supplementary Fig 14. Weighted genomic risk scores (wGRS). a) ROC curves for all Asians constructed by weighted genetic risk scores before (blue), replicated (green) and after (red) adding newly identified SLE loci. b) Density plot of weighted genetic risk scores in Asians. Distribution of wGRS is shown for cases (red) and controls (blue).

Supplementary Fig 15. Principal components for all three populations. First three principal components of our three cohorts KR, HC and MC; together with HapMap/1000 Genomes populations (YRI, CEU, JPT, CHB).

Supplementary Fig 16. Quantile-quantile (Q-Q) plots for SLE associations in datasets from (a) KR, (b) HC, (c) MC, (d) KR+3669 KR controls**.**

Supplementary Table 1. Samples used in this study (IC: ImmunoChip). We used both a Discovery and Replication sets. KR= Korean; HC=Han Chinese; MC=Malaya Chinese.

Supplementary Table 2. ImmunoChip real genotyping SNPs used in this study. Details from the quality control (QC) of the genotyped ImmunoChip SNPs are provided below. MAF= minor allele frequency. HWE= Hardy-Weinberg equilibrium.

Supplementary Table 3. 578 imputation regions with P <5x10-3. IC= Immunochip. KR= Korean. HC= Han Chinese. MC= Malaysian Chinese.

Supplementary Table 4. Six novel regions with suggestive association (Pmeta<5x10-5). We used mach2dat for the single SNP post-imputation-based association tests with the first three principal components as covariates to correct for population stratification in KR, HC and MC, respectively. Imputed SNPs are marked as "*". A1/A2= Minor/major allele. FA/FU= affected/control minor allele frequency. OR= odds ratio. CI= confidence interval. Sample sizes for each cohort are presented as cases/controls. KR=Korean; HC=Han Chinese; MC=Malaya Chinese; JPT=Japanese; SHC=Shanghai Han Chinese; BHC=Beijing Han Chinese.

Supplementary Table 5. All associated SNPs all novel regions, note that *GTF2IRD1-GTF2I* and *SYNGR1* used three ImmunoChip data sets only. FA/FU= affected/control minor allele frequency. OR= odds ratio. CI= confidence interval. Rsq= Imputation quality measure.

Supplementary Table 6. A. Experimental genotyping association results for confirming *GTF2IRD1-GTF2I*. BP= position. A1/A2= minor/major allele. FA/FU= affected/control minor allele frequency. OR= odds ratio. CI= confidence interval. Sample sizes are presented under each cohort as case/control B.Conditional analysis of four *GTF2IRD1-GTF2I* SNPs based on partial genotyping data. C: The linkage disequilibrium (LD) r2 value based on each cohort's control data set.

Supplementary Table 7. Functional annotation of associated lead SNPs, based on Haploreg, ChromMM, and GWAS3D. GERP RS scores of conservation are presented. ChromMM prediction for chromatin state predictions are presented below. LeadSNP=novel SNP used for reference; r2= linkage disequilibrium frequency correlation between the reference SNP and SNP within the feature; SNP =SNP in the feature; TFBS affinity= transcription factor binding site; CTCF = presence of CTCF mark; GERP =conservation status as extracted from GERP; SELF= same SNP as the lead SNP.

Supplementary Table 8. Areas of chromatin interaction. Both Hi-C and CHiA-PET chromatin interaction experimental data from nine different cell lines were used. LeadSNP=novel SNP used for reference; r2= linkage disequilibrium frequency correlation between the reference SNP and SNP within the feature; SNP =SNP in the feature; Locus 1/2= Genomic coordinates of Sides 1 and 2 of the chromatin interaction; Gene1/2= Genes within locus1/2; Score=Read count of interactions; Validation =experiment type.

Supplementary Table 9. Allelic expression changes through eQTLs. eQTLs for both the lead SNP and highly correlated (r2>0.7) SNPs. Probe information presented as chromosome and the center position. Type of eQTL (cis/trans). The gene affected is presented as well as the Source of the data: Bloodeqtl (Westra et al. 2013) and SNPexp(Holm et al. 2010).

Supplementary Table 10. Replicated 36 known loci with P<0.005. KR= Korean; HC =Han Chinese; MC=Malaya Chinese; A1/A2= minor/major allele. FA/FU= affected/control minor allele frequency. OR= odds ratio. CI= confidence interval. Sample sizes are presented under each cohort as case/control. Each SNP was either genotyped in the Immunochip or was imputed.

Supplementary Table 11. Secondary associations for Non-HLA loci. A1/A2= Minor/major allele. FA/FU= affected/control minor allele frequency. OR= odds ratio. CI= confidence interval. Sample sizes are presented under each cohort as case/control. Alleles: I= Insertion D=Deletion.

Supplementary Table 12. HLA-association with classical allele imputation. HLA association was based on dosage data, average case and control dosages are presented. OR= odds-ratio, 95%CI= 95% confidence interval. INFO= imputation quality. For the omnibus test, D is the difference between the base model likelihood and the model with the residue effect.

Supplementary Table 13. All associated SNPs within 6 novel suggestive regions (KRIC+GWAS). A1/A2= Minor/major allele. FA/FU= affected/control minor allele frequency. OR= odds ratio. CI= confidence interval. Sample sizes are presented under each cohort as case/control. Rsq imputation quality.

Supplementary Table 14. Reported SLE SNPs in our cohort. Source of the original report is presented as well as its P-value for each of our cohorts, compared to the European reported p-value and OR. We labeled column "same direction" as NO if the effect of the SNP is opposite in our cohorts, and NA if there was no information on the direction of the effect in the original publication. Source publications are detailed at the right of the table. A1/A2= Minor/major allele. FA/FU= affected/control minor allele frequency. OR= odds ratio. CI= confidence interval. Sample sizes are presented under each cohort as case/control.

Supplementary Table 15. Four loci for which we found highly significant novel uncorrelated SNPs compared to the previously reported GWAS peak SNPs, thereby shifting the association peak. A1/A2= Minor/major allele. FA/FU= affected/control minor allele frequency. OR= odds ratio. CI= confidence interval. Sample sizes are presented under each cohort as case/control. Red color indicated our novel SNPs from this study, black indicated published SNPs. FA=allele 1 frequency of SLE affected; FU=allele1 frequency of unaffected

Supplementary Table 16. Annotation of Bayesian Credible Sets. We used the Bayesian credible set method to better localize the possible set of functional SNPs within each novel region. BF=Bayes Factor; Posterior probability of the SNP assessed. We created a 95% and 99% probability credible set, labeled ad 95 and 99. Each Lead SNP (Table 1 SNP) was labeled as 1. Functional characteristics for each member of the credible set such as protein binding sites, Motif binding, DNAse, Promoter and Enhancer activity, as well as conservation score as determined by GERP/SiPhy were extracted from Haploreg.

Supplementary Table 17. Annotation of the PICS identified SNPs. In order to identify functional SNPs related to our lead SNPs, we used PICS method. The PICS methods assigned a probability of the SNP to be functional within their Linked SNP counterparts (PICS_probability), Functional characteristics for each member of the PICS set such as protein binding sites, Motif binding, DNAse, Promoter and Enhancer activity, as well as conservation score as determined by GERP/SiPhy were extracted from Haploreg.

Supplementary Table 18. Enrichment of drug targets through gene set enrichment analysis. Enriched terms contain= the term, species surveyed, GEO platform accession number, GEO dataset accession number. Overlap: Overlap of our gene set versus the whole set in each category. A combined score for the significance of each category is the combination of the Fisher's P-value with the Z-score estimated from the deviation from the expected mean. For ease of search, we included a column that indicates if any of the novel loci is included in the enriched category.

Supplementary Table 19. Enrichment analysis of Mouse Phenotypes (MGI), as well as GO- Processes. Overlap: Overlap of our gene set versus the whole set in each category. A combined score for the significance of each category is the combination of the Fisher's P-value with the Z-score estimated from the deviation from the expected mean. For ease of search, we included a column that indicates if any of the novel loci is included in the enriched category.

Supplementary Table 20. LD neighborhood for each independent variant. We searched the correlated neighborhood of each Lead SNP for other possible candidate functional SNPs. Ref/alt= reference/alternate alleles. AFR/AMR/ASN/EUR= African, American (from Mexico), Asian, European minor allele frequency.

Supplementary Table 21. SLE-associated regions r2 >0.8 enriched in active chromatin and DNase hypersensitivity sites. Two separate boxes are presented. (a) Enhancer enrichment analysis. (b) DNAse enrichment analysis. Obs: observed value; Exp: expected value; Fold: fold enrichment.

Supplementary Table 22. Summary of natural selection statistics in 1 MB region for 10 novel loci from both the HapMap and HGDP populations. iHS= integrated haplotype score; Xp-EHH= extended haplotype homozigosity.

Supplementary Table 23. Association of genome-wide significant loci with other autoimmune diseases. We searched for reported in which any of our novel or suggestive loci might be associated with any of 12 autoimmune diseases. All citations as well as the reported SNPs are presented below.

Supplementary Table 24. Percentage of heritability explained. a Individual locus heritability; b explained heritability as a whole. RAF= risk allele frequency. Vg= variability. VG Individual significant OR = OR used for the estimation of heritability. VG Individual significant OR= Variance estimated using the cohorts' OR. Heritability explained is presented as the percentage of the reported h2 heritability.

Supplementary Table 25. Fraction of the familial risk (FRR). Two tables are presented. a. SNP based analysis. b. Summary by category of association. For each cohort: Risk= Risk allele frequency; RR= risk ratio. FRR= Korean (KR), Han Chinese (HC), and Malaysian Chinese (MC), assuming sibling risk ratio λs =30.

Supplementary Table 26. Quality controls of the associated SNPs: Association between two controls (IC vs. GWAS), genotyping status (Imputed or real) for the associated SNP, Imputation quality metrics (Rsq) for KR controls (3167) and IC controls (3669). F_IC_control= A1 Allele Frequency in IC controls; F_GWAS=A1 Allele Frequency in GWAS controls.

Supplementary Table 27. Association results for the Japanese Replication(891 cases/3384 controls): SNPs with LD r2>0.6 with the top SNPs within novel genes except *GTF2I*; The SNPs in bold are listed in Table 1 and Supplementary Table 4.

Supplementary Table 28. Association results for the 22 genotyped SNPs on Sequenom array for two independent Han Chinese cohorts from Shanghai (SHC) and Beijing (BHC). Sample sizes are presented as case/control next to each cohort name.

Supplementary Table 29. Over-representation analysis comparison of 100 gene sets taken at random from the Immunochip. We counted each time a pathway/Ontology that was found to be over-represented in our study gene-set appeared in one of the random sets (COUNTinSIM). Size= Size of the gene-set in each category; effective size=corrected size gene set.

Supplementary Fig 1. Venn diagram for 578 imputation regions with P<5x10^{-3} from three populations' ImmunoChip genotyping. KR=Korean; HC= Han Chinese; MC=Malaya Chinese. Numbers represent the number of regions of the genome with association P <5x10^{-3} prior to imputation.

Supplementary Fig 2. Statistical power for combined cases and controls, before and after addition of out-of-study KR controls. It is of note that our initial statistical power comparison between ImmunoChip-only and ImmunoChip + Korean GWAS data sets showed that for MAF > 0.15, adding controls boosted statistical power by 12-16%. For example, using a cutoff at P ≤ 5x10^{-8} and OR = 1.30, the power increased by 15% (from 0.67 to 0.82) at MAF = 0.30.

P=5.0E-8, OR=1.3

KR_IC(1710/3167)

KR_IC+GWAS_ctrls(1710/6836)

Supplementary Fig 3. Regional association plots for 10 novel SLE loci. Plots of -log10(Discovery P-meta) versus position in megabase pairs. The lead SNP is highlighted in purple, while the linkage disequilibrium of the rest of the variants (r2) to the lead SNP is shown in the color legend. a. TERT; b. TCF7; c. IL12B; d. DEF6; e. GTF2I; f. PCNXL3 ; g. RASGRP1; h. CD226; i. SIGLEC6; j. SYNGR1

a.

b.

e.

Supplementary Fig 4. Epigenetic annotation of novel SLE-associated regions. ENCODE tracks for DNase hypersensitivity, HMM promoter/enhancer/insulator/transcription classification, chromatin marks (CTCF, EZH2, H2Az, H3K27Ac, H3K27me3, H3K6me3, H3K4me1, H3K4me2, H3K4me3, H3K79me2, H3K9Ac, H3K9me3, H4K20me1), transcription factor binding sites and RNA-binding protein sites are shown. Data is taken from the ENCODE tracks for GM12878 cells. For chromatin marks and transcription factor/RNA-binding protein sites, shading of the box indicates intensity of the signal, with black meaning maximum signal. The SNPs tested in this study are shown at the top, with the height of the bar (y-axis on left) indicating $-\log_{10}$(P-value). Chromosome position is indicated on the x-axis. For each locus, a blow-up of the region containing the lead SNP and other potentially functional SNPs is shown in a second panel, with key SNPs labeled.

68.8

**SF4a1.
GTF2I/GT
F2IRD1**
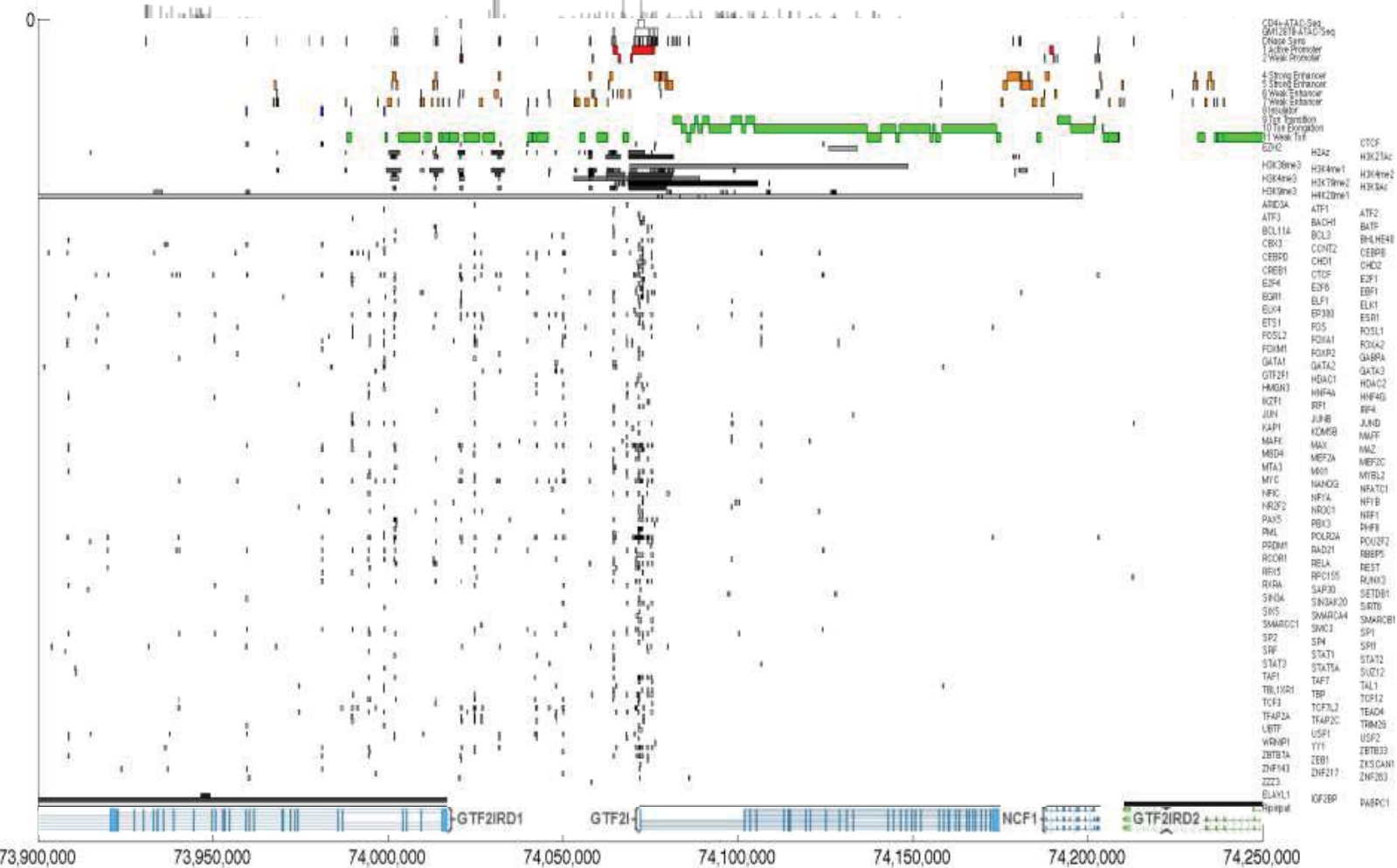
$-\log_{10}(P)$
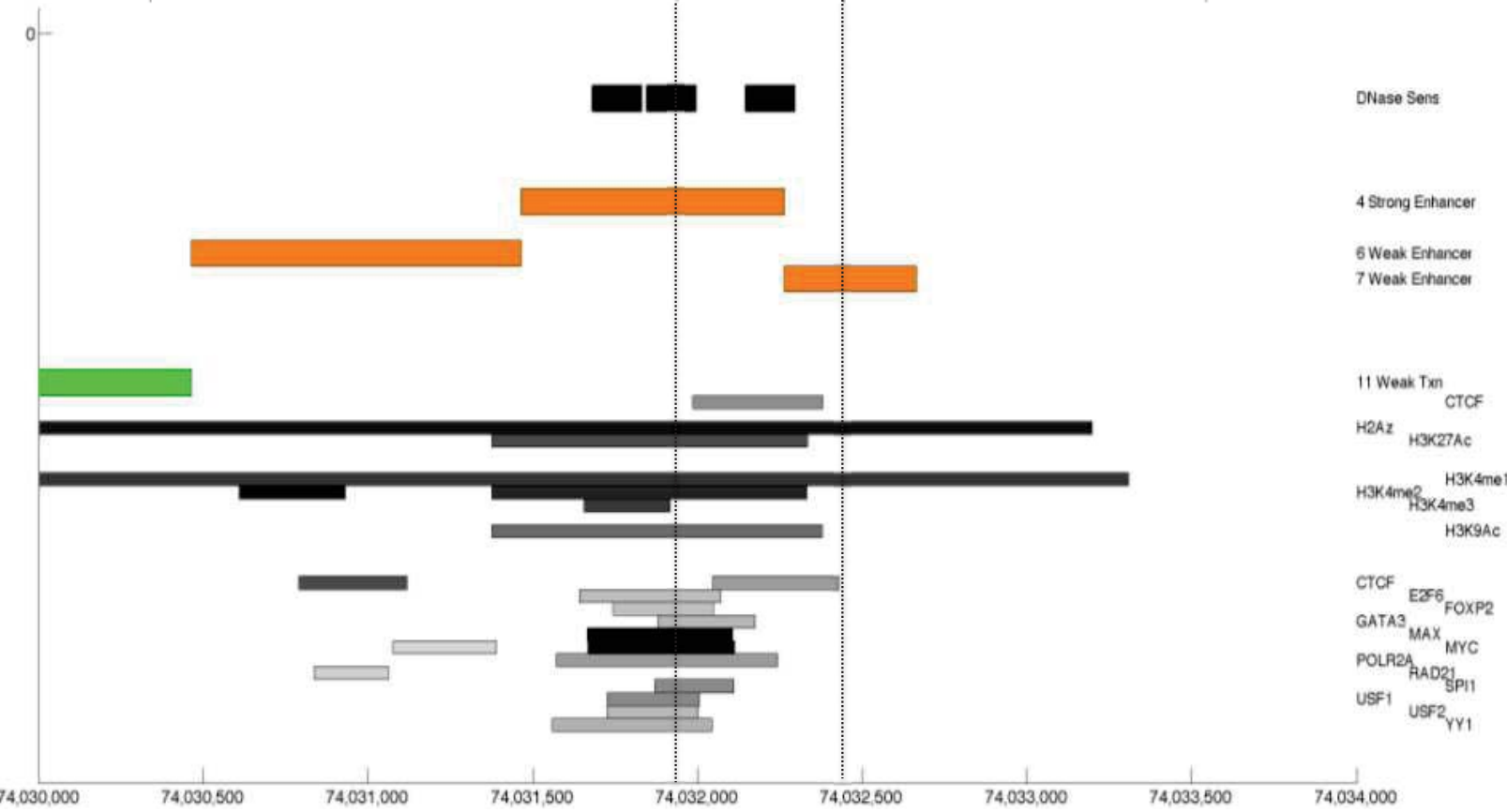
SF4a2. GTF2I/GTF2IRD1

SF4b1. DEF6

SF4b2. DEF6

SF4c1. IL12B

SF4c2. IL12B

SF4d1. TCF7

6.24
-log$_{10}$(P)

0

CTCF
EZH2   HDAc   H3K27Ac
H3K27me3   H3K9me3   H3K4me1
H3K4me2   H3K4me1   H3K9me3
H3K9Ac   H3K9me3   H4K20me1

ARID3A   ATF1   ATF2
ATF3   BACH1   BATF
BCL11A   BCL3   BCLAF1
BHLHE40   BRCA1   BRF2
CBX3   CCNT2   CEBPB
CEBPD   CHD1   CHD2
CREB1   CTBP2   CTCF
CTCFL   EGR1   EGR4
E2F6   EBF1   ESRE1
ELF1   ELK1   ELK4
EP300   ESR1   ESRRA
ETS1   EZH2   FOS
FOSL1   FOSL2   FOXA1
FOXA2   FOXM1   FOXP2
GABPA   GATA1   GATA2
GATA3   GTF2B   GTF3F1
GTF3C2   HDAC1   HDAC2
HMGN3   HNF4A   HNF4G
IKZF1   IRF1   IRF4
JUN   JUNB   JUND
KAP1   KDM5B   MAFF
MAFK   MAX   MAZ
MED4   MEF2A   MEF2C
MTA3   MXI1   MYBL2
MYC   NANOG   NFATC1
NFE2   NFIC   NFYA
NFYB   NR2F2   NR3C1
NRF1   PAX5   PBX3
PHF8   PML   POLR2A
POU2F2   PRDM1   RAD21
RBBP5   RCOR1   RELA
REST   RFX5   RUNX3
RXRA   SAP30   SETDB1
SIN3A   SIN3AK20 SIRT6
SMARCA4   SMARCB1 SMARCC1
SMARCC2   SMC3   SP1
SP4   SPI1   SREBP1
SRF   STAT1   STAT3
STAT5A   TAF1   TAF7
TAL1   TBL1XR1 TBP
TCF12   TCF3   TCF7L2
TEAD4   TRAP2A TRAP2C
THAP1   TRIM28 UBTF
USP1   USF2   WRNIP1
YY1   ZBTB7A ZEB1
ZKSCAN1   ZNF143 ZNF217
ZNF263
ELAVL1   PABPC1 Rpxinpd

VDAC1   LOC101060024   TCF7   SKP1   PPP2CA   CDKL3

133,300,000   133,350,000   133,400,000   133,450,000   133,500,000   133,550,000   133,600,000

SF4d2. TCF7

SF4d3. TCF7

SF4e1. TERT

SF4e2. TERT

SF4f1. CD226
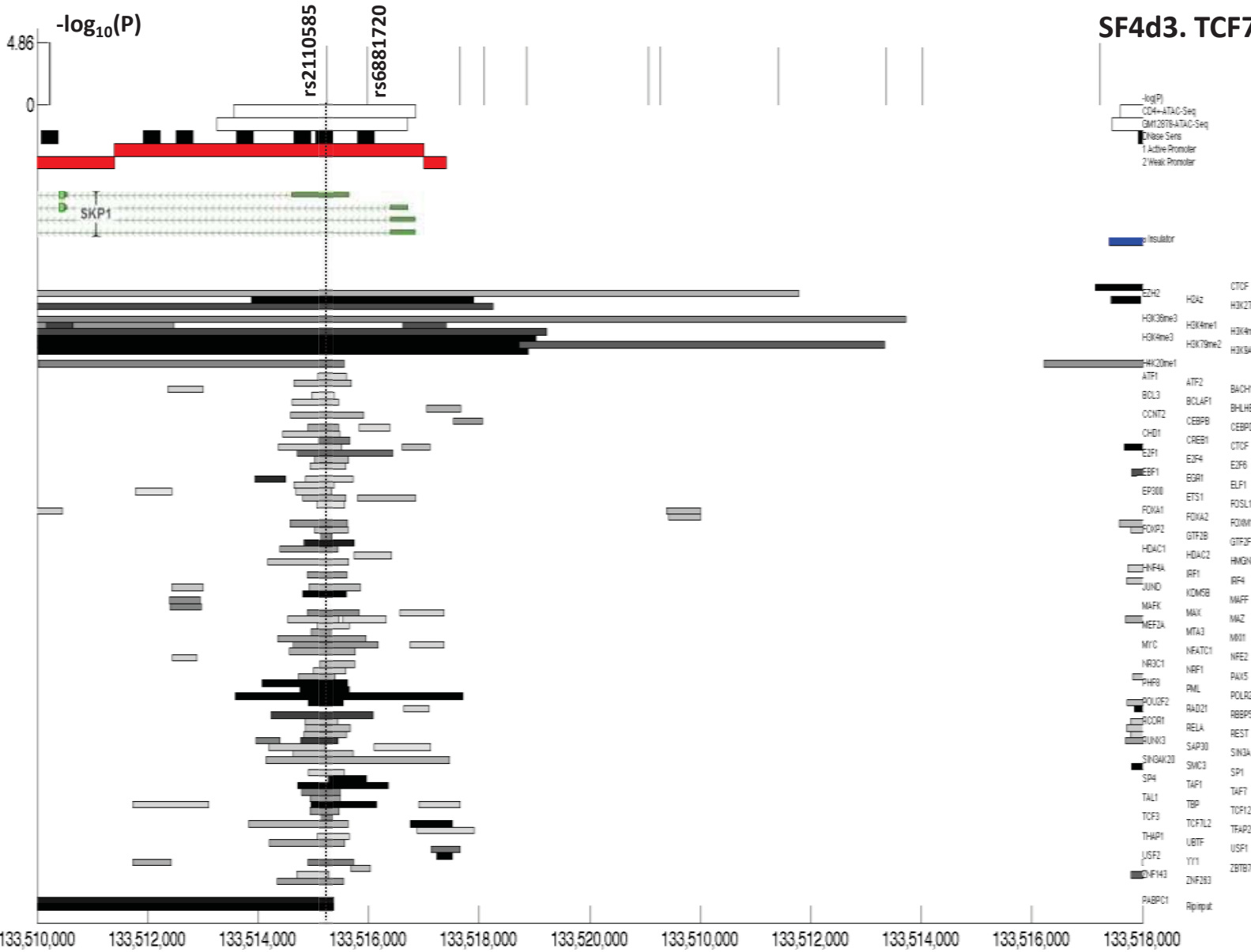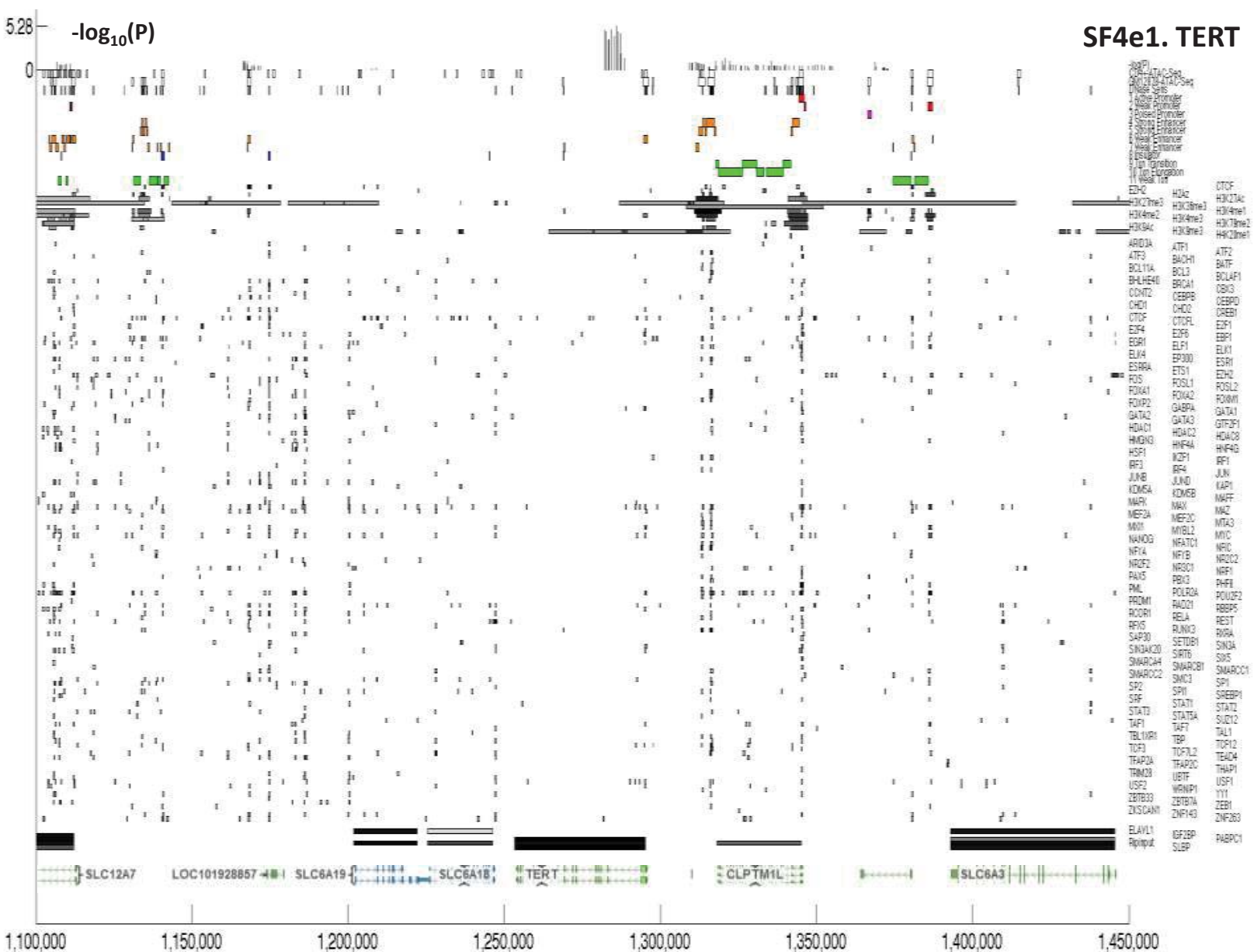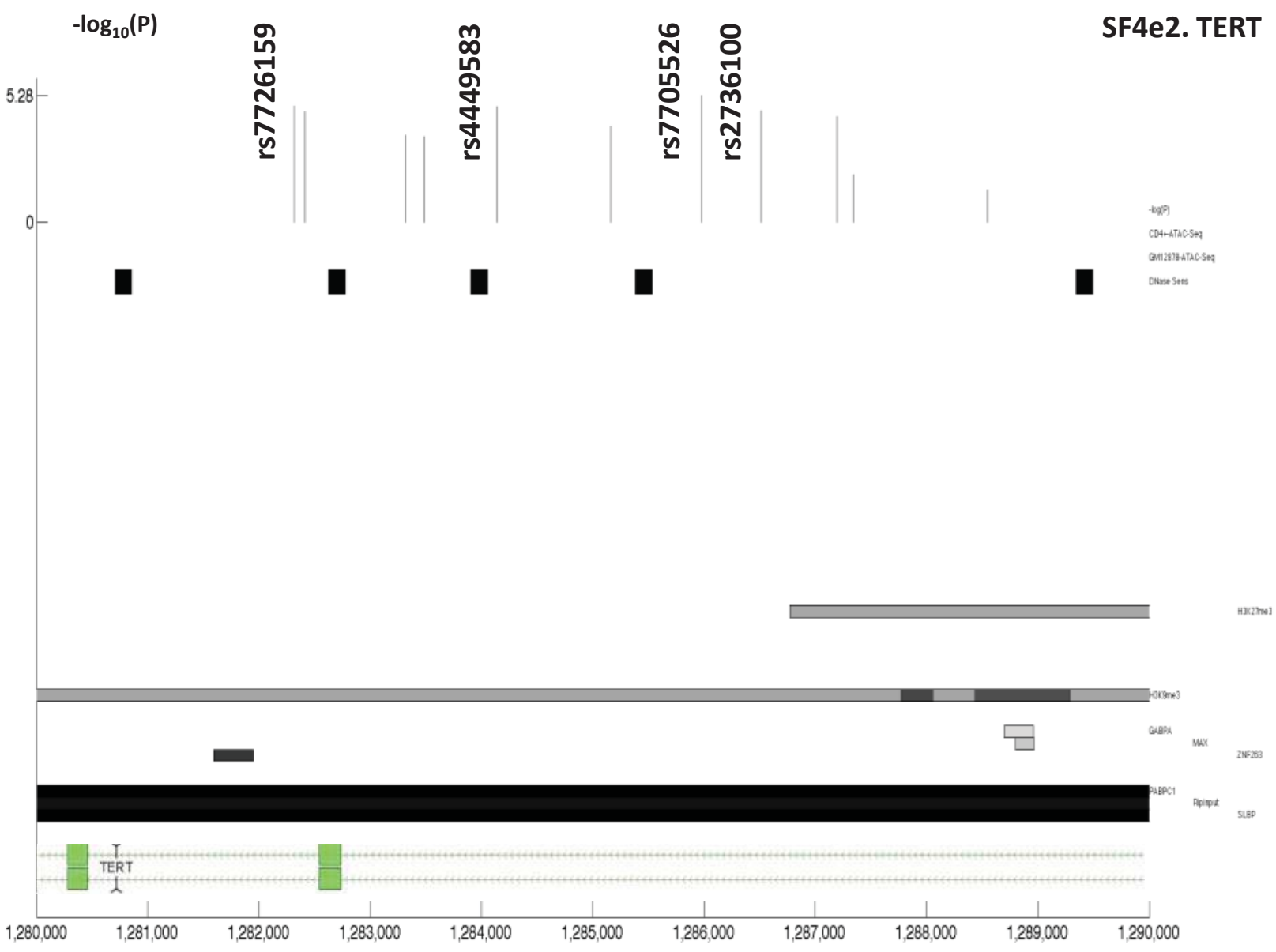
SF4f2. CD226

SF4g1. PCNXL3
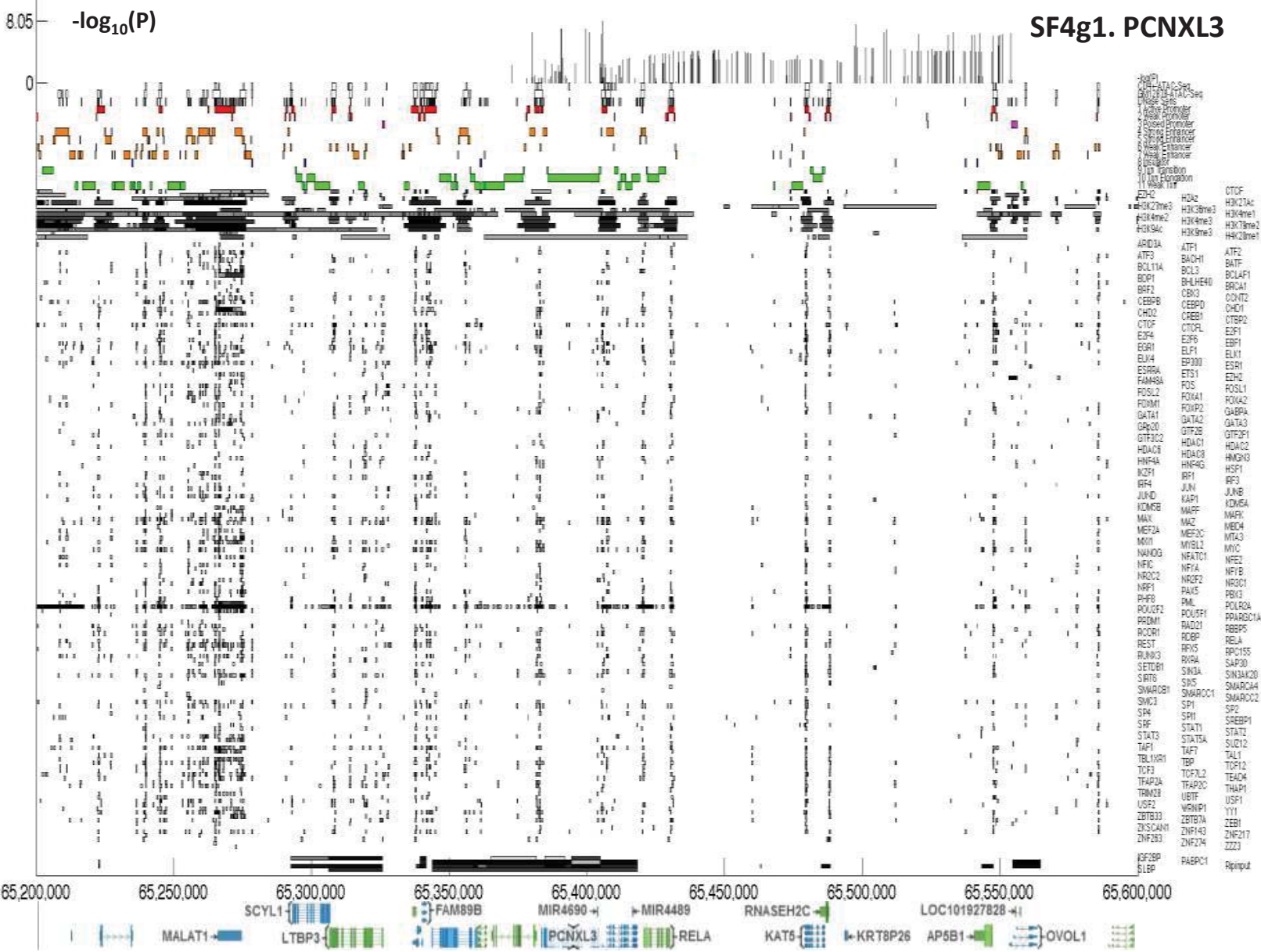
SF4g2. PCNXL3

SF4h1. RASGRP1
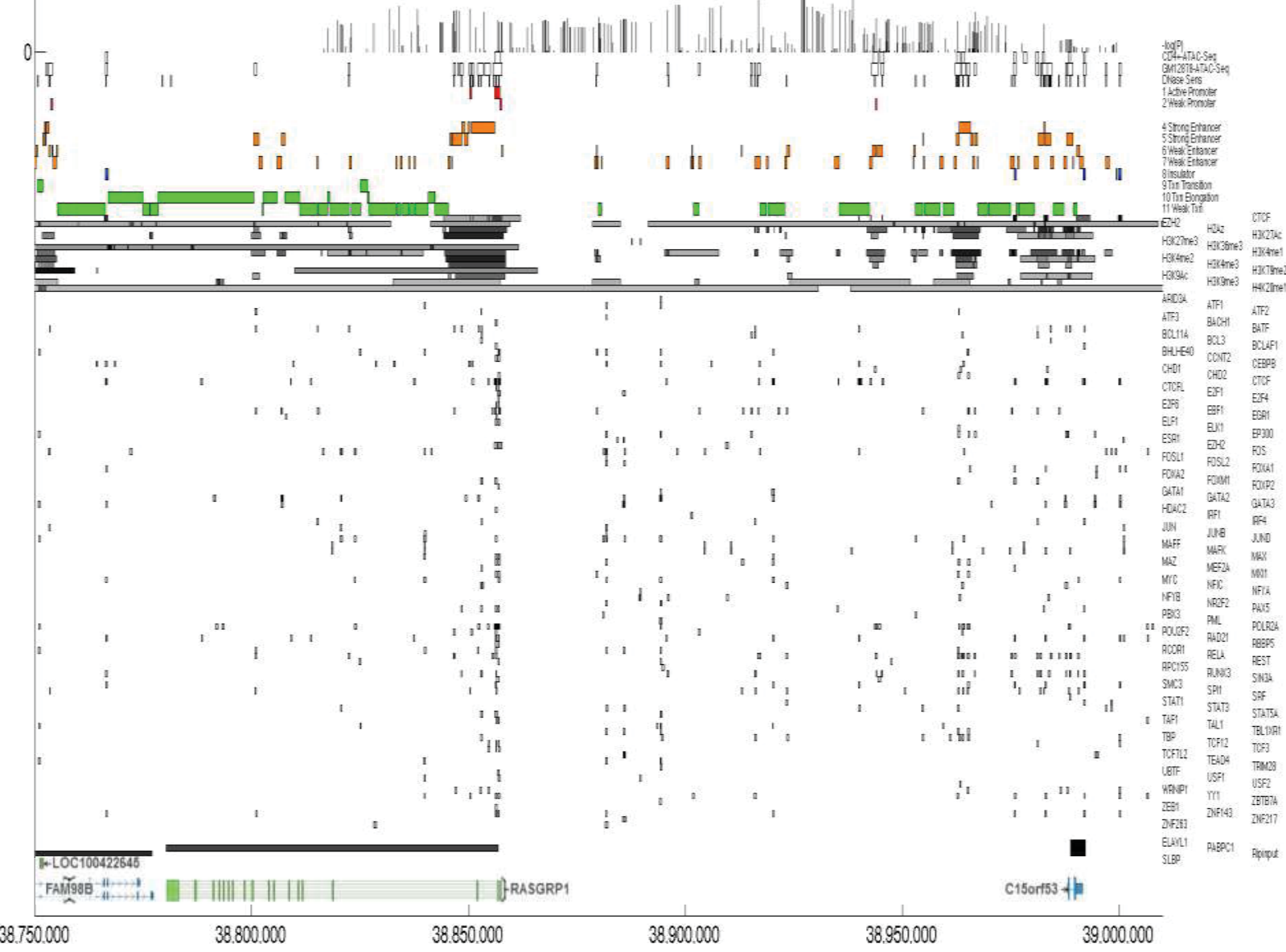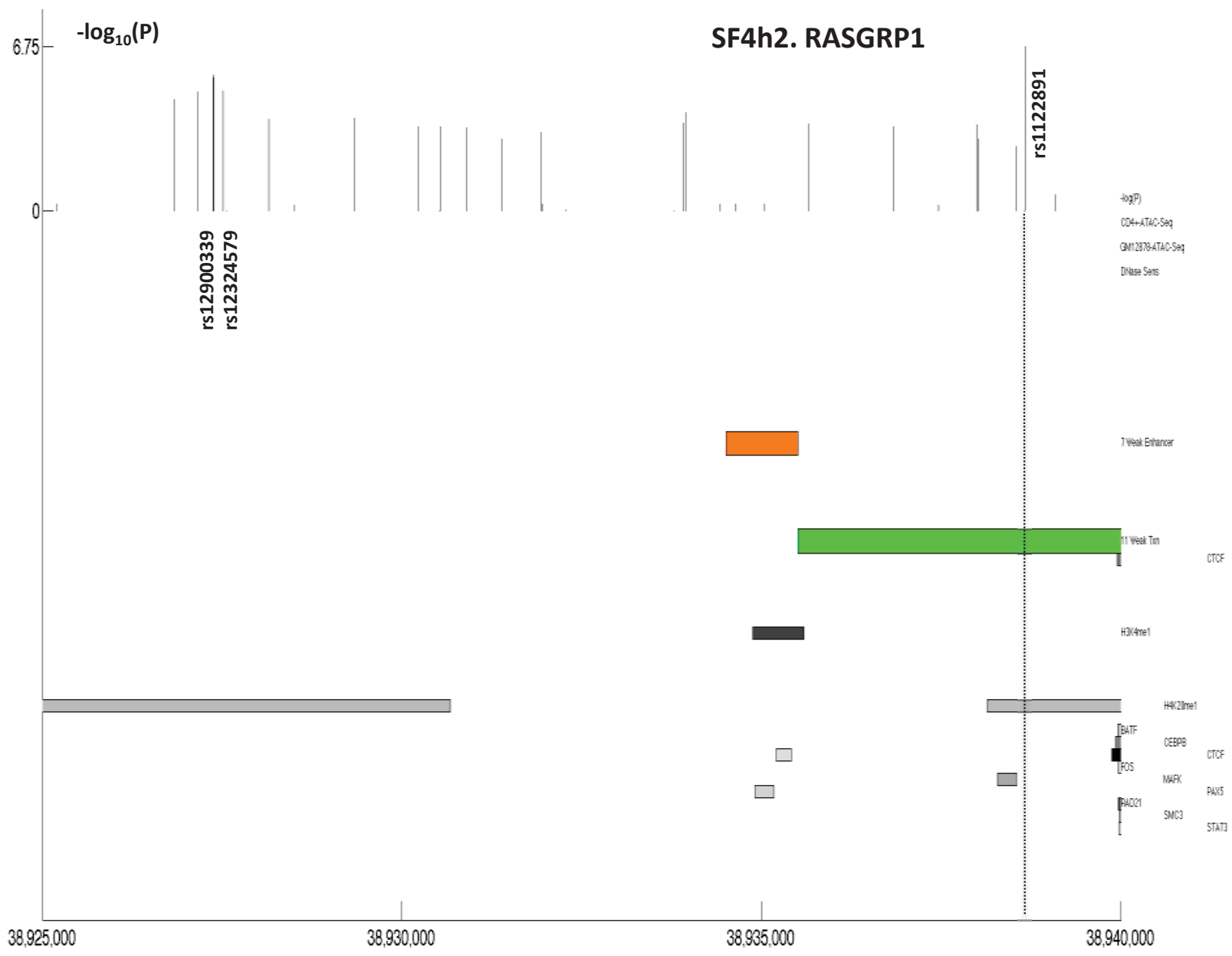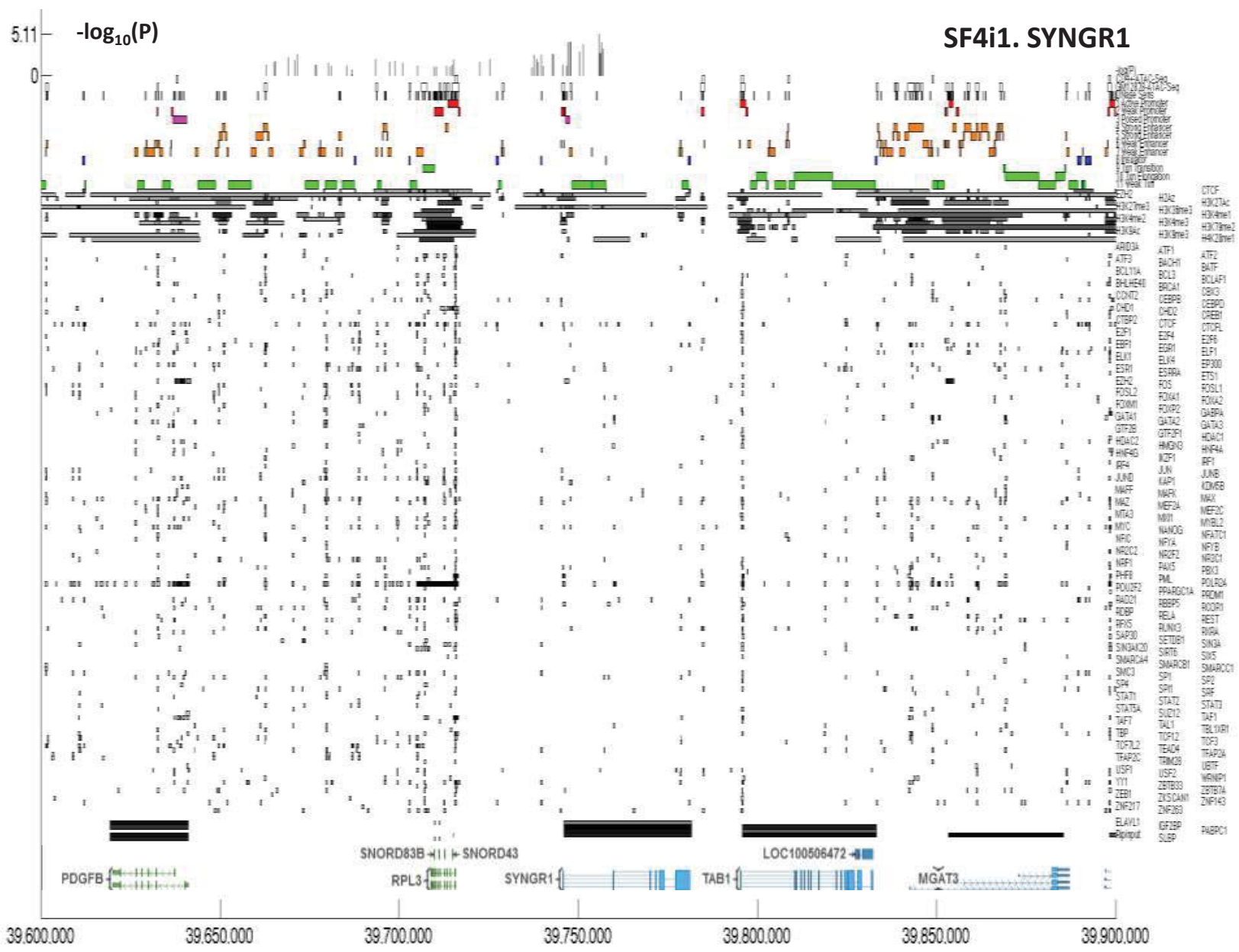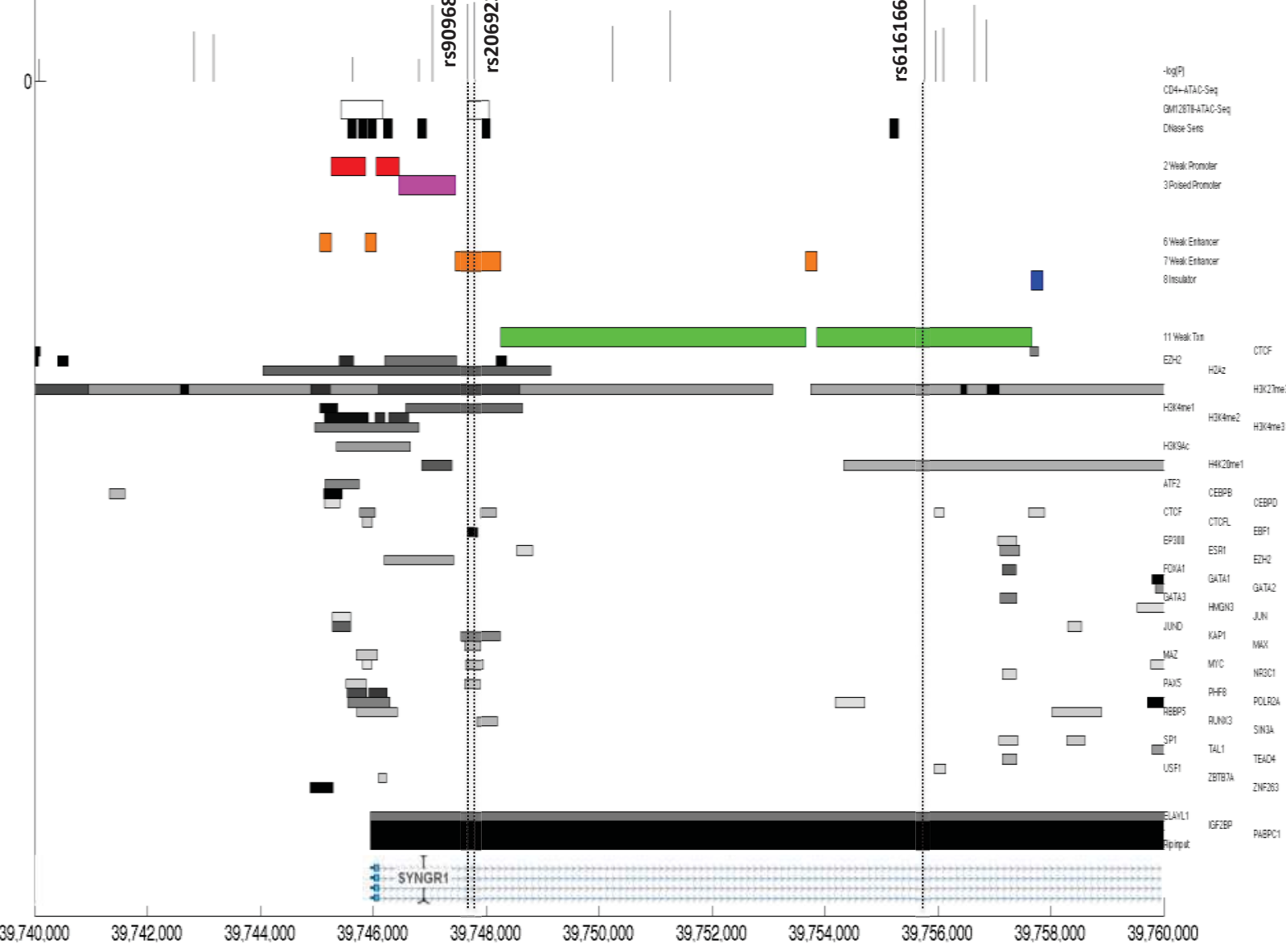
SF4h2. RASGRP1

SF4i1. SYNGR1

$-\log_{10}(P)$

SF4i2. SYNGR1

SF4j1. SIGLEC6

SFj2. SIGLEC6
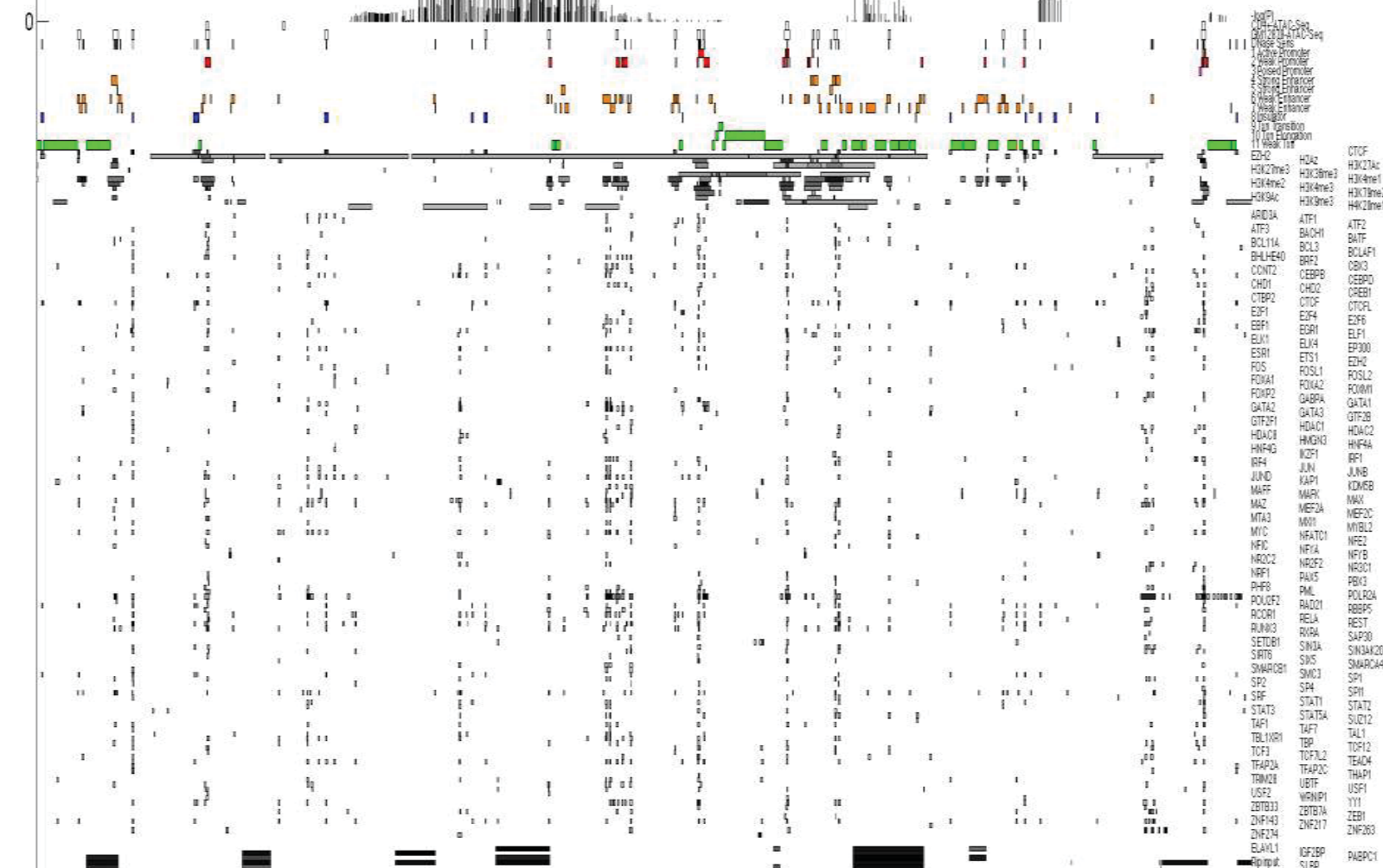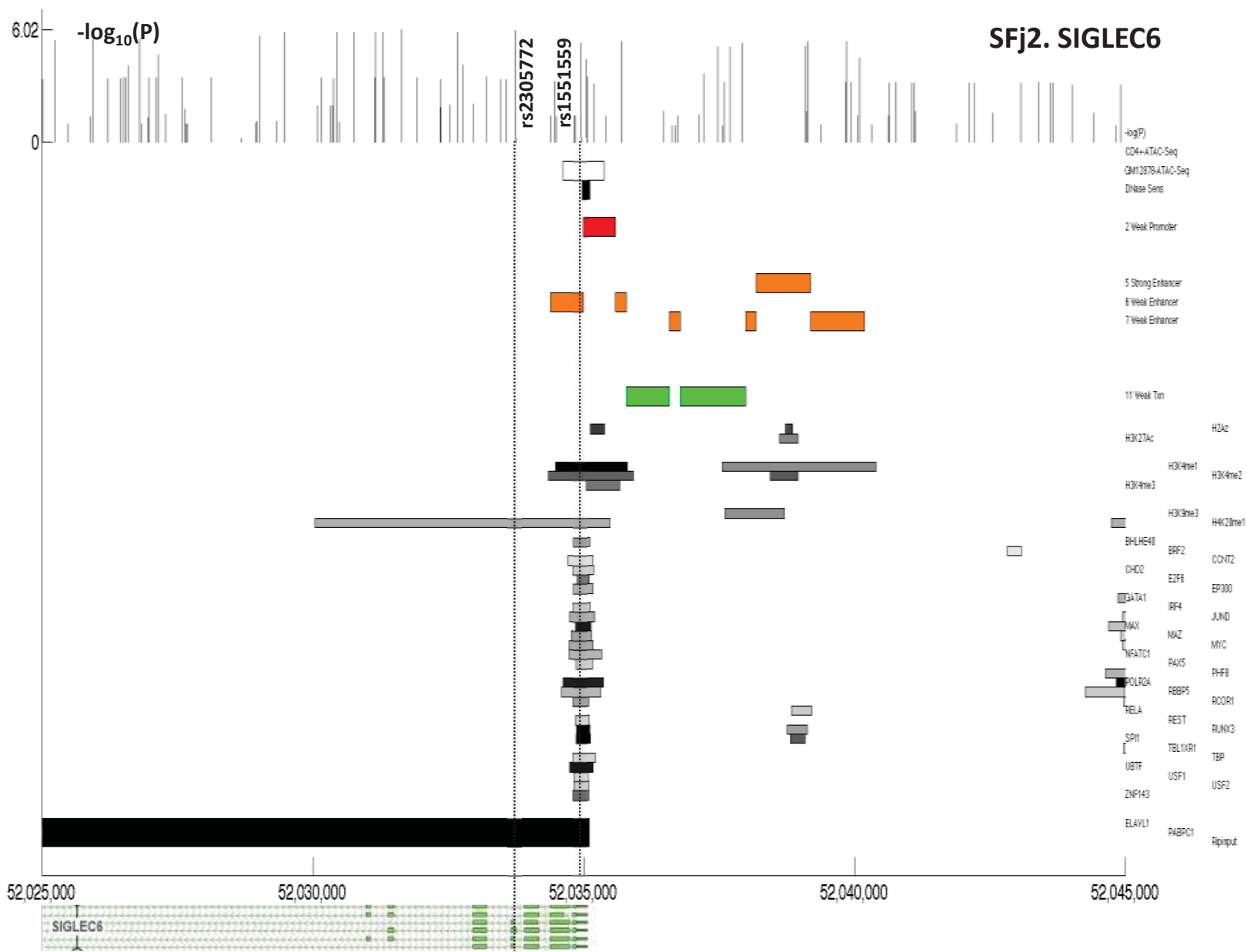
Supplementary Figure 5. Long-range chromatin interactions assessed through Hi-C and ChIA-PET from nine different cell lines. Long range interaction strength is shown in the thickness of the connecting lines.

K562

NB4

GM12878
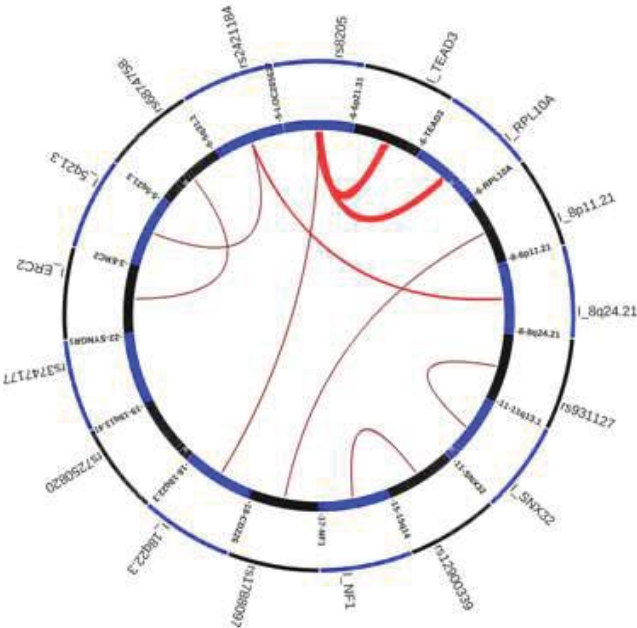
CD4+
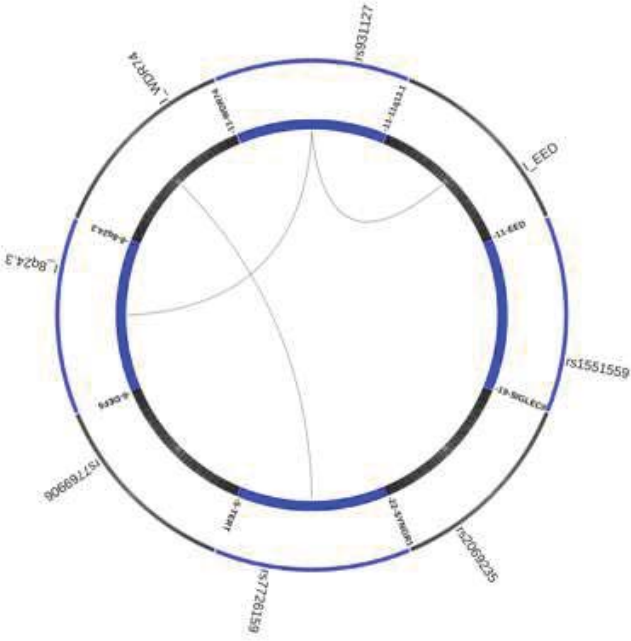
H1-hESC
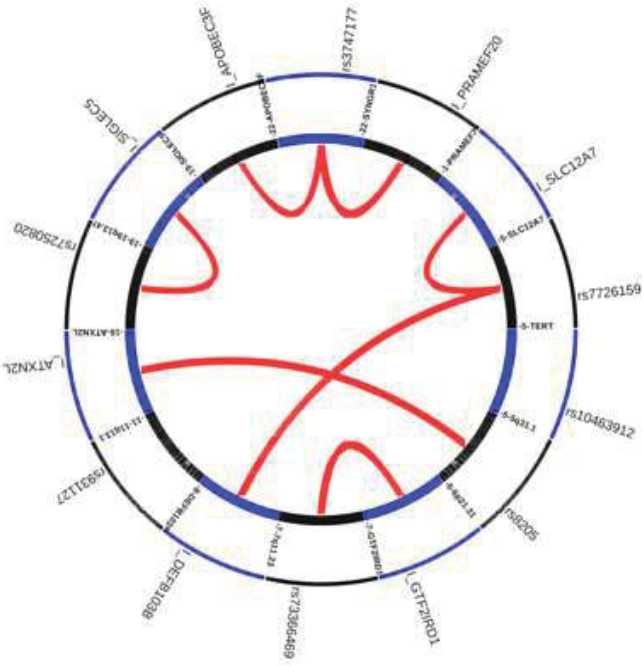
IMR90
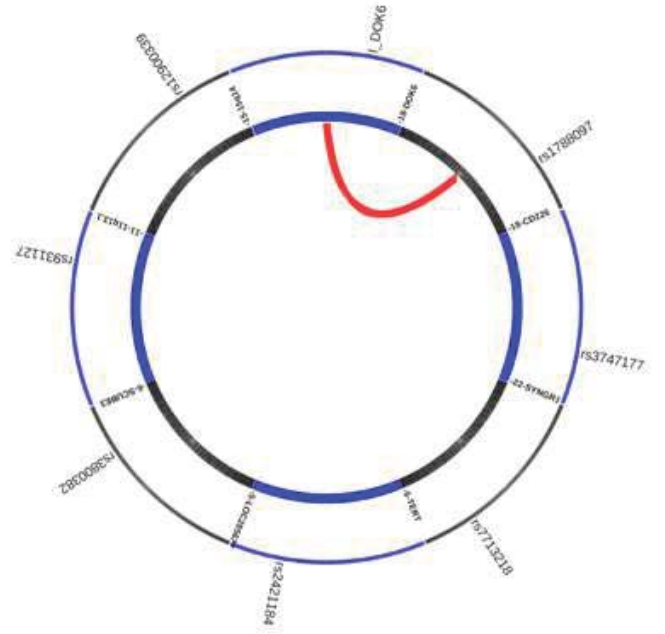
RWPE1
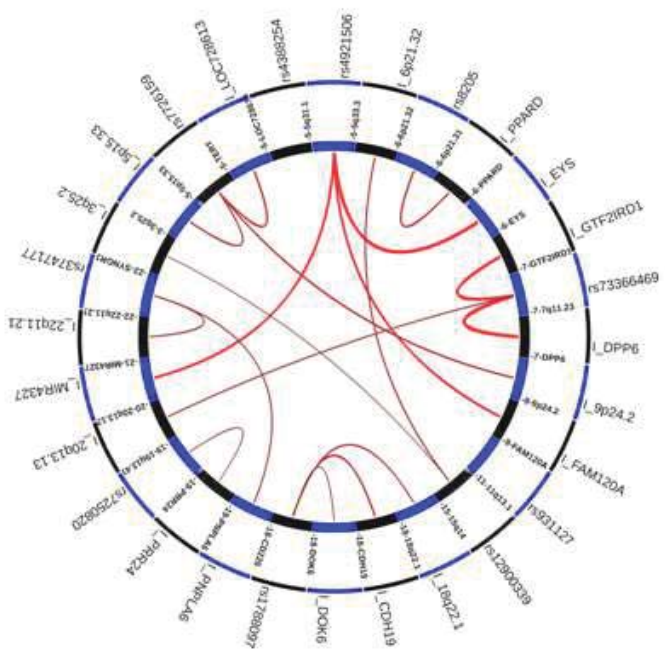
MCF-7
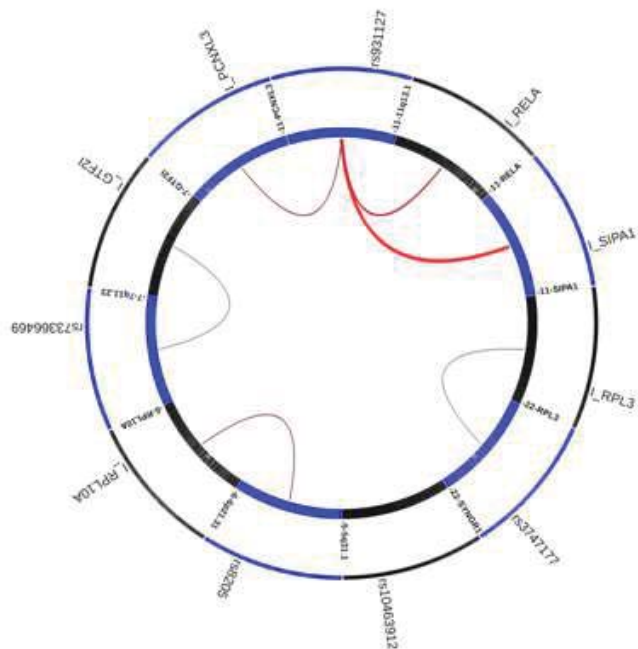
HeLa
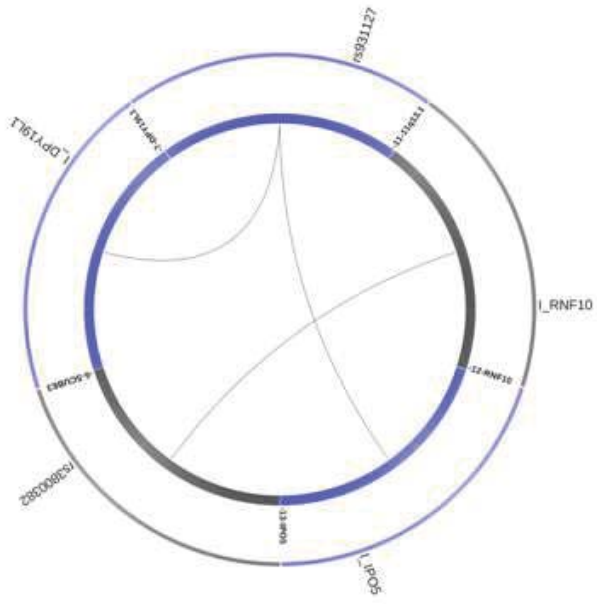
Supplementary Figure 6. Regional association plots for known SLE loci (35 known non-*HLA* genes). Plots of -log10(Discovery P-meta) versus position in megabase pairs. The lead SNP is highlighted in purple, while the linkage of the rest of the variants (r2) to the lead SNP is shown in the color legend. **a STAT4; b ETS1; c TNFAIP3; d TNFSF4; e IRF5; f IKZF1; g BLK; h HIP1; i IRAK1; j WDFY4; k IRF8; l TNIP1; m RASGRP3; n PRDM1_ATG5; o ARHGAP31_TMEM39A_CD80; p SLC15A4; q UBE2L3; r BANK1; s CLEC16A; t PDHX_CD44; u SNRPC; v LBH; w DDX6_CXCR5; x FCGR2B; y PTTG1; z FCRL5; aa BACH2; ab PRKCB; ac NCF2; ad CSK; ae JAZF1; af PHRF1; ag TLR7; ah UBE2E3; ai IL4**

Supplementary Fig 7. Regional association plot for the extended MHC/HLA region. Plots of -log10(Discovery P-meta) versus position in megabase pairs. The lead SNP is highlighted in purple, while the linkage of the rest of the variants (r2) to the lead SNP is shown in the color legend. a SNP based association analysis; b, Imputed HLA variants (classical allele and residue) association; c, After conditioning for the effect of *HLA-DRB1*, no association remained. We positioned a red dotted line at the genome-wide significant level (P<5x10-8), as well as a suggestive association line (5x10^{-8}<P<5x10^{-5}).

b

C

Supplementary Figure 8. Enrichment analysis of the novel and SLE loci in cell-specific gene ontology categories. We estimated enrichment of our gene-set in the Gene Ontology database. Overrepresented cell types have a high correlation (Pearson's correlation coefficient) of SLE-loci expression (dark red). P-values (blue bars) that passed the multiple testing threshold (black line) show significant enrichment in SLE loci.

Supplementary Fig 9. Induced network module analysis. In order to investigate how our novel loci fit together, we created an induced network connecting our set of novel loci to interacting genes (blue), proteins (yellow) and drug targets (red). Red stars mark novel loci.

Supplementary Figure 10. Protein-protein interactions for each novel SLE locus. a *TERT*; b. *TCF7*; c. *IL12B*; d. *DEF6*; e. *GTF2I*; f. *PCNXL3; g. SIGLEC6*

a

b

c

I: interleukin 12B (natural killer cell stimulatory factor 2, cytotoxic lymphocyte maturation factor 2, p40)

Physical entity color	Interaction node color	Edge color				Edge style	
gene	biochemical reaction	Manual upload	Reactome	Kegg	Humancyc	→	substrate / product
protein	gene regulation	Pid	Biocarta	Netpath	Inoh	⊸●	enzyme
protein complex	physical interaction	Ehmn	Phosphositeplus	Innatedb	Intact	⟹	transcribed gene
RNA	genetic interaction	Dip	Mint	Hprd	Corum	⋯○	activator
compound	drug-target interaction	Biogrid	Mips-mppi	Bind	Pdb	⋯⊣	inhibitor
family / unknown		Matrixdb	Pindb	Spike	Pig	⋯⊙⊣	regulator (unknown/dual)
		Phosphopoint	Pharmgkb	Smpdb	Signalink	⋯	physical interactor
		Drugbank	Ttd	Chembl	Wikipathways		
		Pdzbase					

d

e

Physical entity color
- gene
- protein
- protein complex
- RNA
- compound
- family / unknown

Interaction node color
- biochemical reaction
- gene regulation
- physical interaction
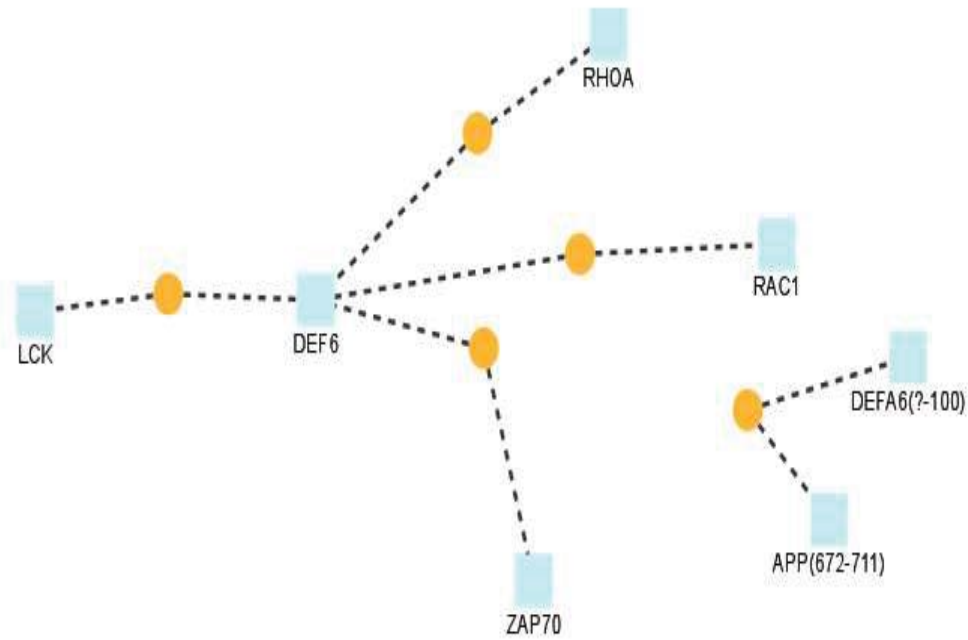- genetic interaction
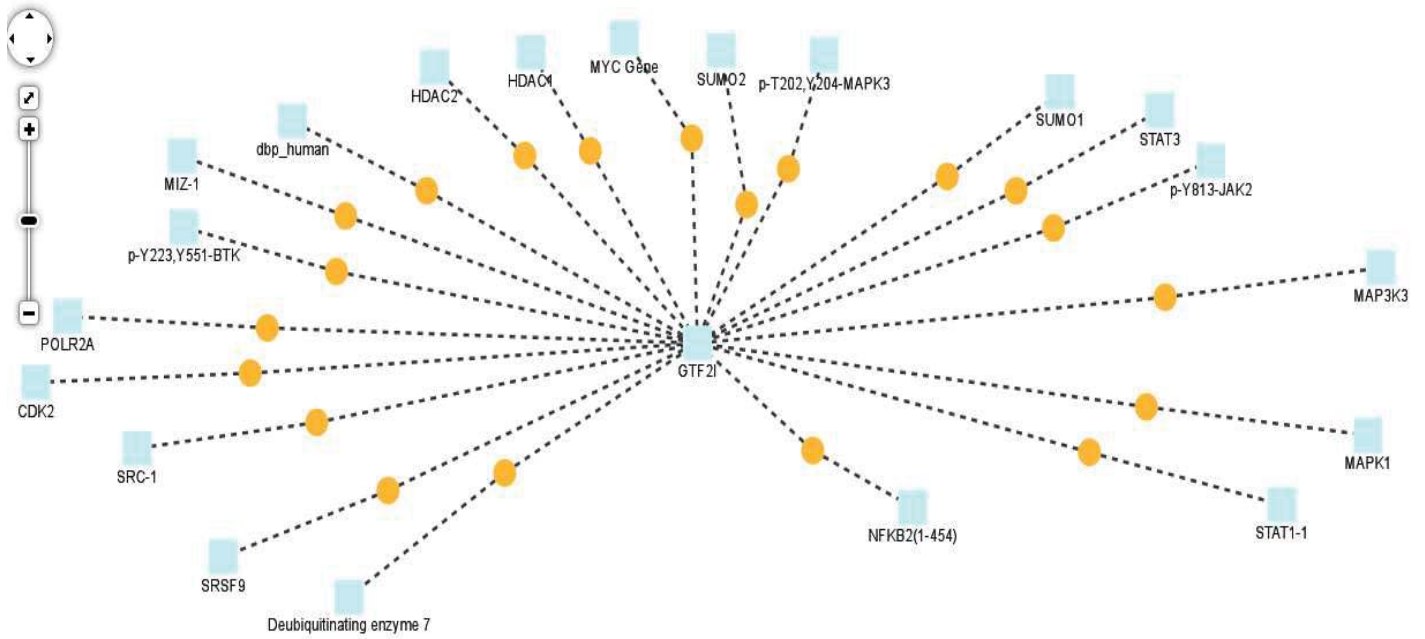- drug-target interaction

Edge color
- Manual upload
- Pid
- Ehmn
- Dip
- Biogrid
- Matrixdb
- Phosphopoint
- Drugbank
- Pdzbase
- Reactome
- Biocarta
- Phosphositeplus
- Mint
- Mips-mppi
- Pindb
- Pharmgkb
- Ttd
- Kegg
- Netpath
- Innatedb
- Hprd
- Bind
- Spike
- Smpdb
- Chembl
- Humancyc
- Inoh
- Intact
- Corum
- Pdb
- Pig
- Signalink
- Wikipathways

Edge style
- substrate / product
- enzyme
- transcribed gene
- activator
- inhibitor
- regulator (unknown/dual)
- physical interactor

f

Q81KM0_BACAN [Bacillus anthracis] PCNXL3 RECR_FRATT [Francisella tularensis subsp. tularensis]

g

Q81SP4_BACAN [Bacillus anthracis] SIGLEC6 Y936_YERPE [Yersinia pestis]

| Physical entity color	Interaction node color	Edge color				Edge style	
gene	biochemical reaction	Manual upload	Reactome	Kegg	Humancyc	→	substrate / product
protein	gene regulation	Pid	Biocarta	Netpath	Inoh	●—	enzyme
protein complex	physical interaction	Ehmn	Phosphositeplus	Innatedb	Intact	▶—	transcribed gene
RNA	genetic interaction	Dip	Mint	Hprd	Corum	····○	activator
compound	drug-target interaction	Biogrid	Mips-mppi	Bind	Pdb	····⊣	inhibitor
family / unknown		Matrixdb	Pindb	Spike	Pig	····○⊣	regulator (unknown/dual)
		Phosphopoint	Pharmgkb	Smpdb	Signalink	········	physical interactor
		Drugbank	Ttd	Chembl	Wikipathways		
		Pdzbase					

Supplementary Figure 11. Network illustrating published relationships between known Genes (blue) and novel Genes (green). Relationships are defined as the co-occurrence of gene names (or their synonyms) within the same abstract. Line thickness is proportional to the number of times the genes were mentioned together. Node size is proportional to the number of connections within this network (i.e., bigger = more connections). Genes that were not co-mentioned with other genes on the list are shown at bottom and divided according to how much published information on them exists (poorly <= 15 papers mentioning the gene in any context, Moderately = between 16 and 50 papers, Well-annotated >=100 papers)
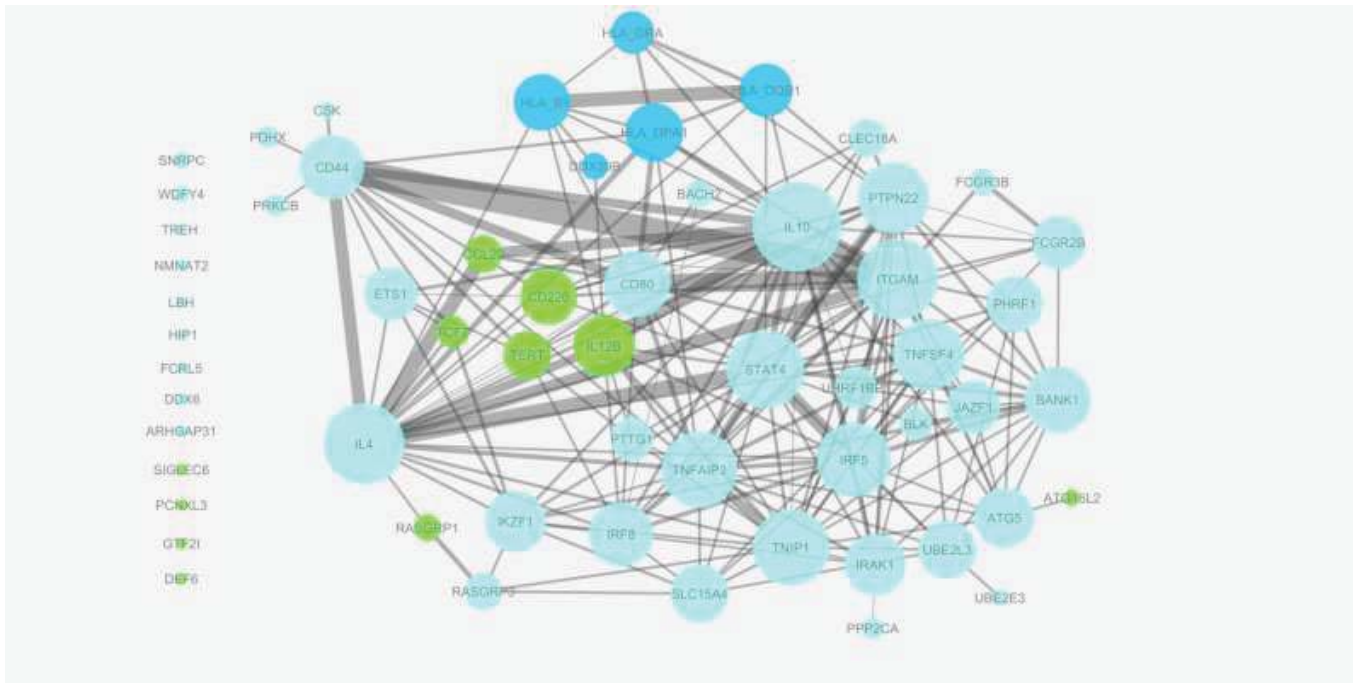
Supplementary Figure 12. Protein-protein interaction sub-groups identified through DAPPLE analysis of reported biochemical interactions.
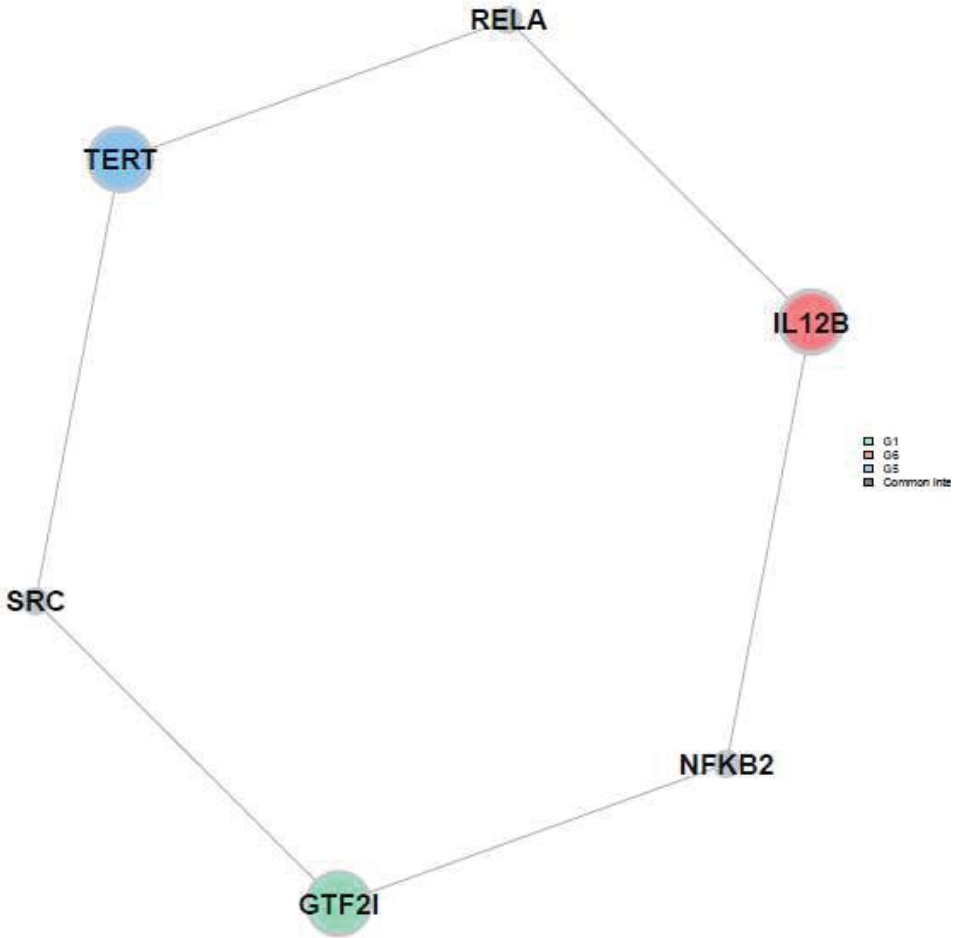
Supplementary Figure 13. Enrichment analysis of the novel and known SLE loci in cell type-specific gene expression. We estimated enrichment of our gene-set in human (a: GeneAtlas) and mouse (b: ImmGen) cell lines. Overrepresented cell types have a high correlation (Pearson's correlation coefficient) of SLE-loci expression (dark red). P-values (blue bars) that passed the multiple testing threshold (black line) show significant enrichment in SLE loci.
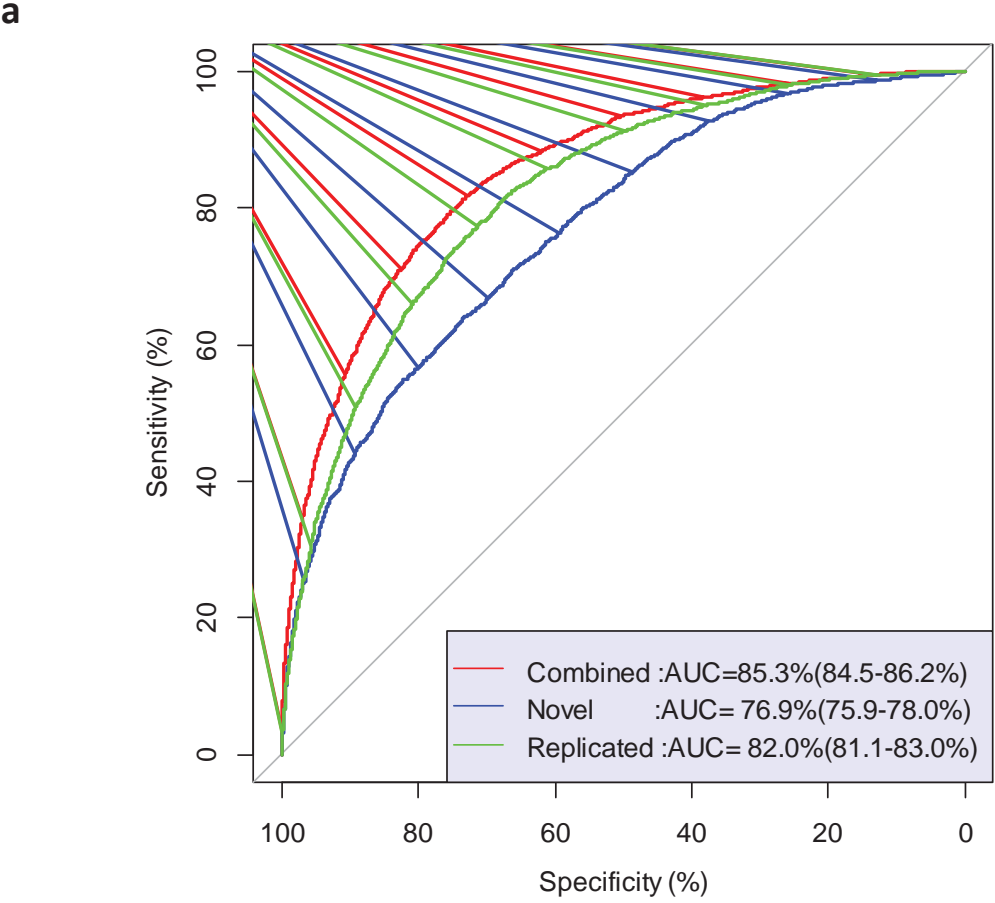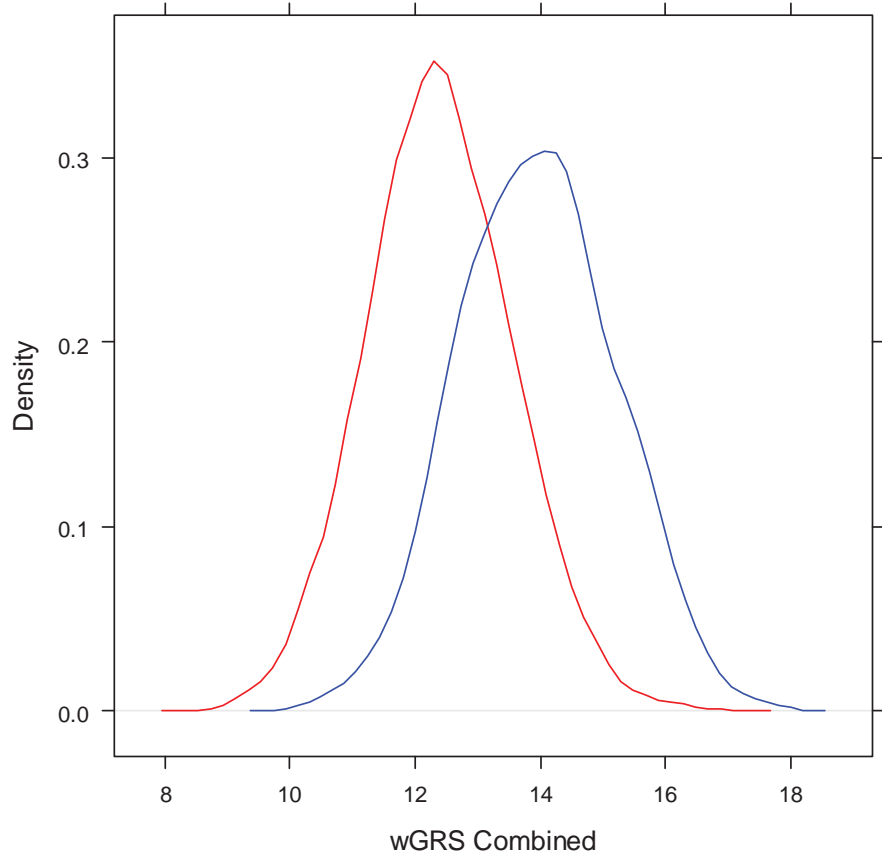
a. GeneAtlas
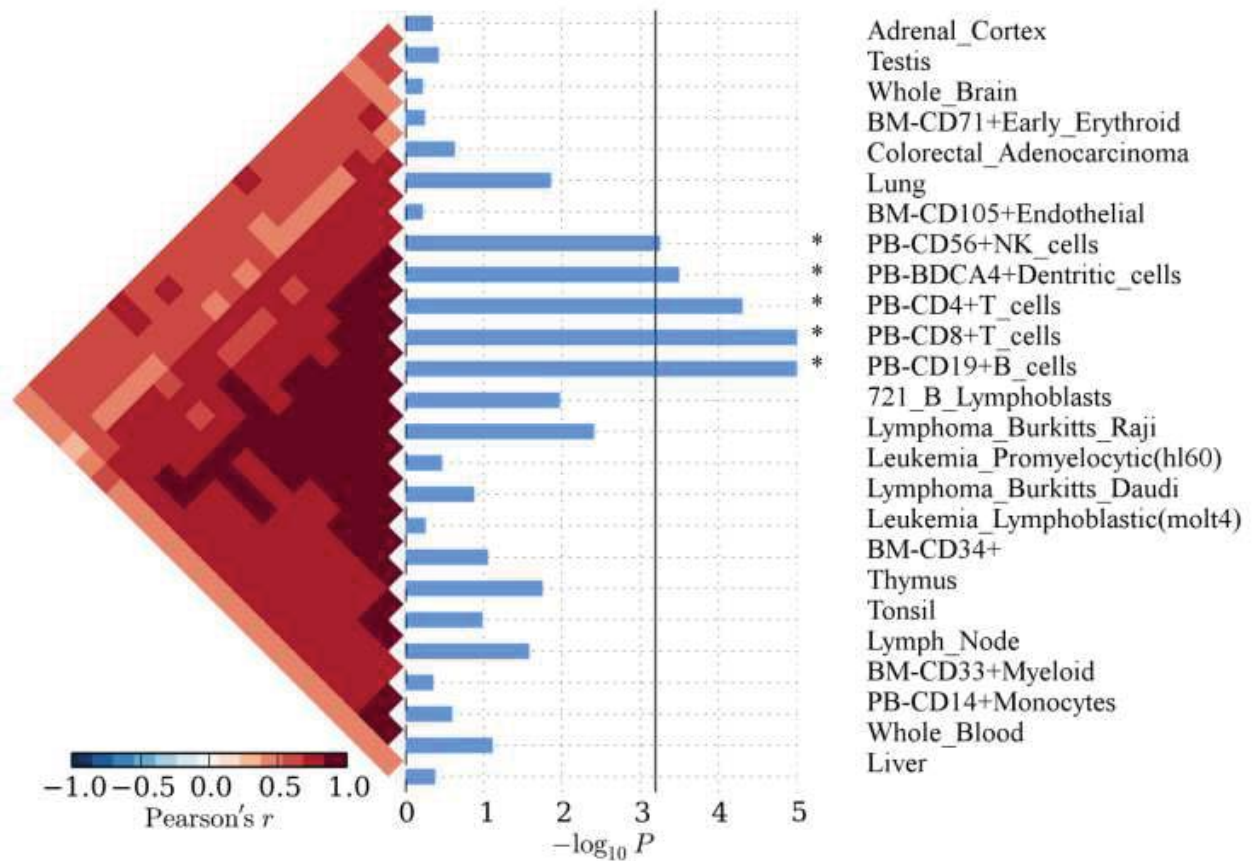
b. ImmGen
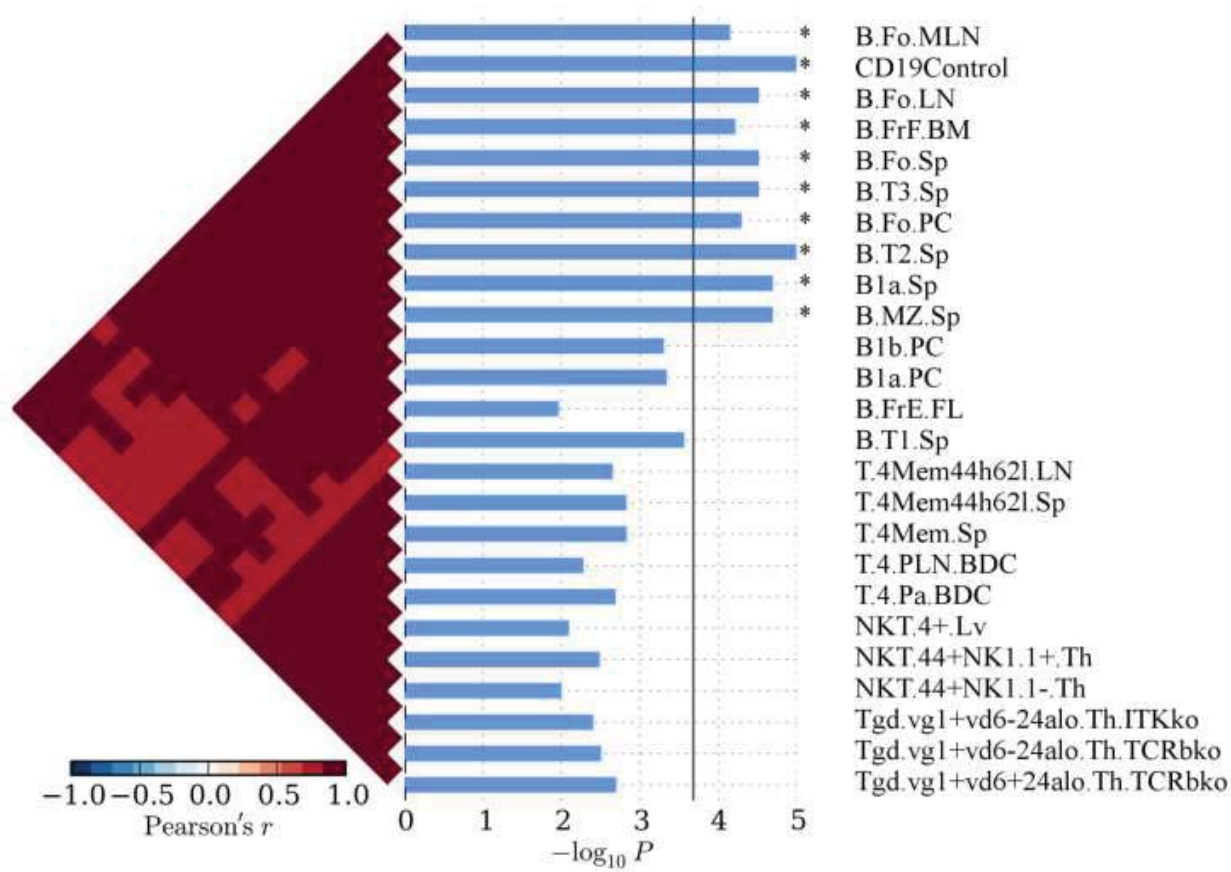
Supplementary Figure 14. Weighted genomic risk scores (wGRS) for 47 loci. ROC curves (a) for all Asians constructed by weighted genetic risk scores Novel loci (blue), replicated SLE loci (green) and total (red). Density plot of weighted genetic risk scores in (b) Asians. Distribution of wGRS is shown for case (red) and control (blue).

b

Supplementary Figure 15. Principal components for all three populations. Three first principal components of our three cohorts KR, HC and MC; together with 1000 genomes populations (YRI, CEU, and ASN =JPT+CHB).
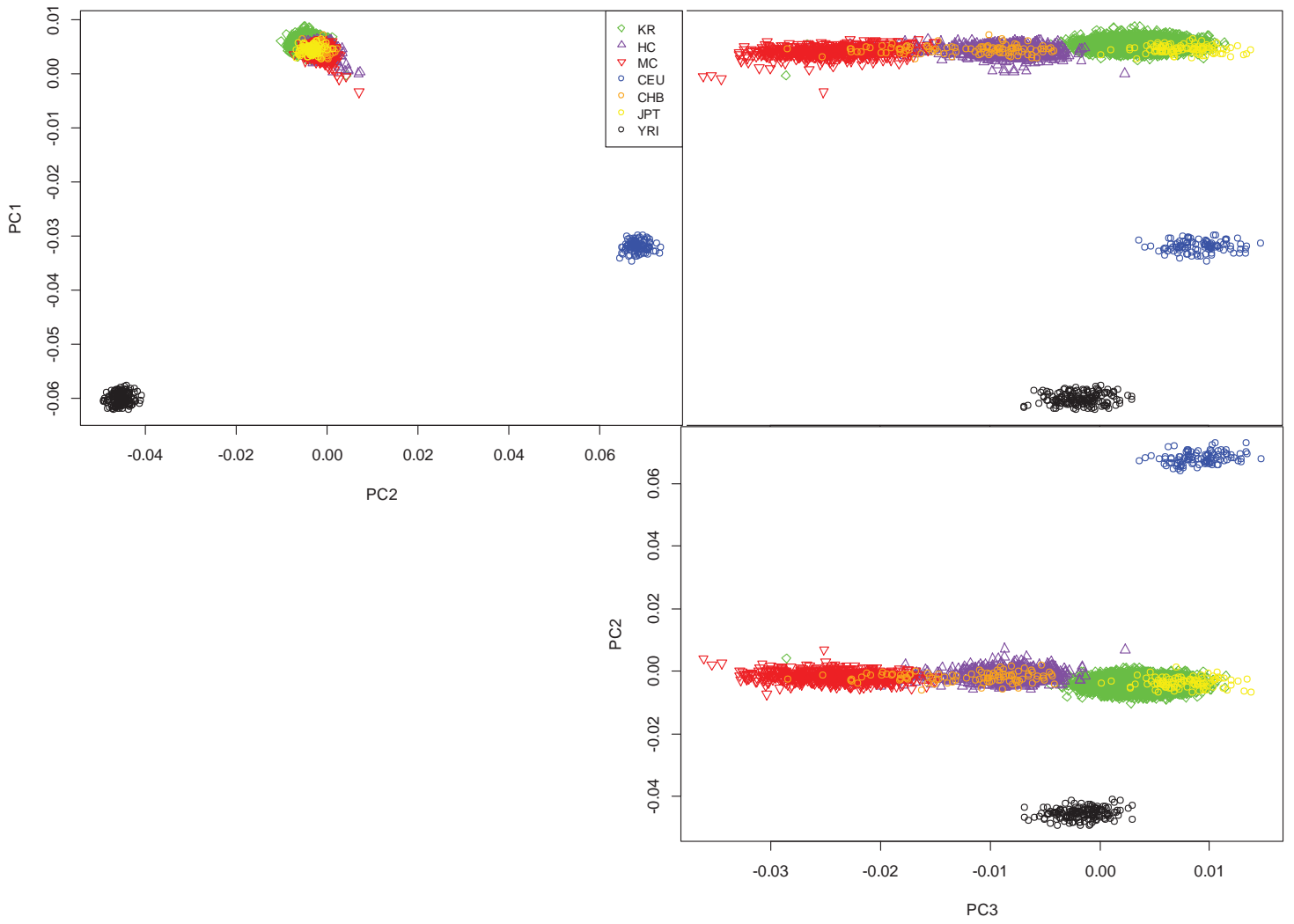
Supplementary Figure 16. Quantile-quantile (Q-Q) plots for SLE associations in datasets from (a) KR, (b) HC, (c) MC, (d) KR+3669 KR controls,