Supplementary methods

Immunophenotyping

To sort CD8 TN ($CD3^+CD8^+CD45RA^+CCR7^+$), CD8 TCM ($CD3^+CD8^+CD45RA^-CCR7^+$), CD8 TEM ($CD3^+CD8^+CD45RA^-CCR7^-$), CD8 TEMRA ($CD3^+CD8^+CD45RA^+CCR7^-$), $HLADR^+$ CD8 T-cells ($CD3^+CD8^+HLADR^+$), CD4 TN ($CD3^+CD4^+CD45RA^+CCR7^+$), CD4 TCM ($CD3^+CD4^+CD45RA^-CCR7^+$), CD4 TEM ($CD3^+CD4^+CD45RA^-CCR7^-$), basophils ($CCR3^+CD123^+$), and regulatory B-cells (BREG) ($CD19^+CD24^+CD38^+$)[1], the following antibodies were used: anti-CD3 (APC/Cy7, BioLegend, UCHT1), anti-CD4 (PE/Cy7, BioLegend, SK3), anti-CD8 (Horizon V500, BD, RPA-T8), anti-CCR7 (Alexa647, BioLegend, TG8/CCR7), anti-CD45RA (BV421, BioLegend, HI100), anti-HLA-DR (PerCP/Cy5.5, BioLegend, L243), anti-CD19 (BV421, BioLegend, HIB19), anti-CD24 (PerCP/Cy5.5, BD, ML5), anti-CD38 (FITC, BioLegend, HIT2), anti-CD123 (BV421, BioLegend, 6H6), and anti-CCR3 (APC/Cy7, BioLegend, 5E8).

Transcriptome data preparation

For the Affymetrix human genome U133 plus 2.0 array, probe set expression values were estimated using frozen RMA, and their presence call was calculated with the MAS5 algorithm. Stepwise quality control for probes was conducted. First, probes that did not match any genes or that targeted multiple genes were removed. Then, probes that were called absent in more than one-third of samples in all the comparison groups, i.e., HC and pSS, were filtered out. In the case where multiple probes hybridized the same gene and those probes were positively correlated with a Pearson's correlation coefficient of more than 0.3, the probe exhibiting the maximum average signals across samples was kept. Lastly, less variable probes whose interquartile ranges were in the bottom 20% of all probes were filtered out. After the application of these quality control steps, 12,231 probes (10,187 genes), 15280 probes (10,678 genes), 13909 probes (10,479 genes), and 14,230 probes (12,106 genes) remained for the whole-blood transcriptome, the CD8 T-cell transcriptome, CD4 T-cell transriptome, and the salivary gland transcriptome (GSE23117), respectively.

Genotype data preparation

Genotype data for human immune cell subsets (accession codes EGAD00010000144 and EGAD00010000520) were downloaded from the European Genome-phenome Archive. The most recent annotation file for the Illumina OmniExpress v1.0 chip was obtained from Illumina, Inc. as of 2015. Samples and SNP quality control steps were carried out with PLINK v1.07. Genotype data of EGAD00010000144 and EGAD00010000520 were merged after adjusting for strand flipping of variants and removing variants whose strands were undetermined (AT or GC). Redundant markers of genomic position were collapsed by keeping markers with the lowest missing rates. We sequentially removed individuals who (i) were involved in pairs of related individuals closer than second-degree relatives as detected by the proportion of identity by descent (PI_HAT>0.25) (n=5), (ii) with call rates<95% (n=1), and (iii) with autosomal heterozygosity more than 3 standard deviations away from the global mean (n=7). After individual quality control steps were applied, SNPs (i) with MAF<1%, (ii) with missing rates>10%, and (iii) that deviated from Hardy-Weinberg equilibrium (p-value<1e-50 as recommended by PLINK) were filtered out. Finally, 646,575 SNPs from 419 individuals remained and were used for imputation.

After the chromosome was chuncked into fragments of 2550 bp with 500 bp overhangs, each fragment was phased using the mach software with 25 iterations of Markov sampling for 300 haplotypes and a random drawing of haplotypes every five iterations. The phased genotypes were subjected to imputation by Minimac3 based on the reference panel of 1000 Genomes Phase 3 populations. SNPs with an estimated rsq greater than 0.3 and a genotype call probability greater than 0.95 were considered as usable for dosage data and hard-called genotype data, respectively. In total, 13,347,774 SNPs were measured or imputed, and both were then combined as dosage data.

Collecting eQTL information

Expression data of human immune cell subsets (accession codes E-MTAB-3536, E- MTAB-2232, and E-MTAB-945) were downloaded from ArrayExpress. Probe annotation data of Illumina HumanHT12v4 chip were obtained from the illuminaHumanv4.db R package. We defined a probe detection call as present if the detection p-value was less than 0.01. Stepwise quality control for probes was conducted as follows. Probes that were not detectable in all the samples were removed. Probes were further removed if the ProbeQuality was evaluated as Bad or No match in illuminaHumanv4.db. We further filtered out probes that hybridized the genomic

sequences harbouring SNPs with more than 1% of minor allele frequency in the European (EUR) panel of 1000 Genomes Phase 3. Collapsing of probes targeting the same gene and removal of less variable probes were conducted in the same procedure described above. All samples were used if the corresponding genotype data were not dropped during the quality control process. When there were duplicated samples for identical individuals, the samples with the largest number of detectable genes were used. The qualified data that went through all the filtering steps were as follows: 12,468 probes (10,836 genes) from 366 monocytes stimulated with interferon gamma, 12,425 probes (10,516 genes) from 279 monocytes, 12,766 probes (11,075 genes) from 260 monocytes with 2 hours of lipopolysaccharide (LPS) stimulation, 12,142 probes (10,814 genes) from 321 monocytes with 24 hours of LPS stimulation, 13,160 probes (11,188 genes) from 278 B-cells, and 10,708 probes (9514 genes) from 101 neutrophils. Expression data were normalized by removing hidden covariates estimated using the peer method[2]. The number of hidden covariates used for the normalization was determined by visual inspection of the saturation of the number of genes with cis-eQTLs (eGene). eGenes were calculated using SNPs located within 1 Mbp of gene body using FastQTL software[3] with 1000 random permutations. The numbers of hidden covariates selected were as follows: 16 for monocytes stimulated with interferon gamma, monocytes with 2 hours of lipopolysaccharide (LPS) stimulation, and monocytes with 24 hours of lipopolysaccharide (LPS) stimulation; 10 for monocytes; 12 for B-cells; and 2 for neutrophils. After removing the hidden covariates from the expression matrix, the eQTLs statistic for each SNP was calculated using Matrix eQTL[4]. Large-scale blood eQTL data[5] and eQTL results for sorted immune cells[6] were obtained from the supplementary table of each report. eQTL data was filtered out at the threshold of p-value less than 0.05 and FDR less than 0.05.

pSS GWAS enrichment

We obtained 9 SNPs that reached genome-wide and suggestive significance levels in GWAS for pSS in Han Chinese population[7]. The SNPs in a linkage disequilibrium with the pSS GWAS SNPs were estimated based on 1000 Genomes Phase 3 data from East Asian population via the r2 function in PLINK with ld-window as 99999, and ld-window-kb as 1000. The nearby genes of the pSS GWAS SNPs were defined as the genes whose coding regions are overlapped with the proximal GWAS SNPs whose R2 is greater than 0.5. Alternatively, GWAS SNPs were assigned to genes by combining the eQTL list with the proximal GWAS SNPs whose R2 is greater than 0.8. The enrichment of the pSS GWAS gene sets in the omics modules was evaluated using the

fisher's exact test.

Cell count imputation

Surrogate variables that correspond to the amounts of CD4 T-cells, CD8 T-cells, NK cells, B-cells, plasmablasts, neutrophils, monocytes, and eosinophils were estimated based on whole-blood gene expression data. Probes that were highly correlated with immune cell amounts relative to white blood cells (p-value<0.05, q-value<0.05, Pearson's correlation coefficient>0.7) were defined as the cell signature probes using samples (n=48 or 49) for which both transcriptome and cell count information were available (Supplementary Figure 2a, Supplementary Table 2). The cell specificity of the probe expression in corresponding cell types was confirmed with the use of expression profiles of purified immune cell subsets from IRIS (GSE22886)[8] and DMAP[9] (Supplementary Figure 2b,c). Normalized expression data of IRIS and DMAP were obtained from the CellCODE R package[10]. To estimate the relative number of immune cells, principal component analysis under the non-negativity constraint (the nsprcomp R package) was applied to z-scaled expression data of the cell signature probes (Supplementary Figure 3a,b).

Differential correlation testing

We used a standardized z-score-based differential correlation test[11, 12]. Specifically, Pearson's or Spearman's correlation coefficient in each condition was transformed to a z-score using Fisher's transformation. The z-score approximately follows a normal distribution; therefore, the difference in z-scores between two conditions is also approximately normally distributed. The p-value of differential z-scores was then enumerated against the null standard normal distribution.

The differential correlation of gene modules was assessed based on the enrichment of gene pairs differentially correlated between healthy and pSS. Specifically, for each module, the proportion of differentially correlated links was compared with that of other modules using Fisher's exact test.

Gene set enrichment analysis

The significance of the overlapping of two gene sets was assessed with Fisher's exact test. MSigDB hallmark gene set collection[13] and canonical pathways from the IPA (Ingenuity Systems, www.ingenuity.com) were used. For the enrichment with MSigDB, the Enrichment Map was utilized for visualization of the results[14]. The gene set variation analysis (GSVA) was used

for pathway enrichment analysis for CD8 T-cells transcriptome[15]

Differentially methylated regions enrichment

Differentially methylated regions in whole-blood samples from pSS were obtained from the supplemental materials provided by Imgenberg-Kreuz *et al* [16]. Genome coordinates of Illumina's 450k methylation arrays were updated from hg18 to hg19 using the IlluminaHumanMethylation450kanno.ilmn12.hg19 R package. The GREAT algorithm[17] was used for the enrichment analysis of cis-regulatory regions with genes of interest. Because the GREAT server provided by the authors does not allow user-defined gene sets, we implemented the algorithm internally. The BSgenome.Hsapiens.UCSC.hg19 and TxDb.Hsapiens.UCSC.hg19.knownGene R package were used for background information on genomes and genes, respectively. The genomic region was assigned to the two nearest upstream and downstream genes whose transcriptional start sites were located within 1000 kb of the region. As described in the literature[17], a binomial test over the genomic regions was performed using binom.test in R given user-defined genomic regions and gene sets.

Differential expression analysis

Identification of transcripts or proteins differentially expressed between pSS and HC was conducted based on the empirical Bayes method using the limma R package. The RNA integrity number and the surrogate variables for the relative numbers of CD4 T-cells, CD8 T-cells, NK cells, B-cells, plasmablasts, neutrophils, monocytes, and eosinophils were used for covariates of the linear model. The false discovery rate was controlled based on q-values estimated with the qvalue R package. We set the criterion for statistical significance at p-value<0.05 and q-value<0.25.

Module identification

Coexpression networks of whole-blood transcriptomes were built and clustered using the WGCNA R package[18]. The topological overlap matrix was generated using pSS transcriptome data with unsigned biweight midcorrelations with a soft thresholding power of three. Then, the dynamic tree cut with the deepSplit parameter of four was applied to hierarchical clustering dendrograms of the topological overlap matrix. Following the package's tutorial, clusters with dissimilarity less than 0.1 were merged.

To cluster serum protein data, an affinity propagation algorithm was employed[19] rather than WGCNA due to the lack of scale-free topology of protein co-abundance networks. Affinity propagation was applied to a correlation matrix of pSS protein data using the apcluster R package with default parameters. Then, the hierarchical relations of clusters were estimated based on exemplar-based agglomerative clustering using the aggExCluster function in the package. According to cluster relationships, similar clusters were merged such that Pearson's correlation coefficients of eigenvalues of every pair of clusters were less than 0.7.

Associations between modules with pSS disease phenotypes were evaluated based on module eigenvalues that were the first principal component whose direction was aligned with the average expression of the module genes. We performed statistical testing based on a linear model using the limma R package with the RNA integrity number as a covariate.

Module preservation

The modulePreservation function in the WGCNA package was used to evaluate preservation of transcriptional modules found in whole-blood transcriptomes. The modulePreservation method takes two types of statistics, density preservation statistics, and connectivity based statistics. The density preservation statistics indicate whether genes in a module are highly correlated each other. The connectivity based statistics assess whether the correlation pattern between genes in the whole-blood data resembles with that in the other data of interest. These statistics were summarized to obtain a composite metric. We performed 100 random permutations to evaluate the statistical significance. The detail of the method is described in[20].

Supplementary references

[1] Blair PA, Noreña LY, Flores-Borja F *et al.* CD19+CD24hiCD38hi B Cells Exhibit Regulatory Capacity in Healthy Individuals but Are Functionally Impaired in Systemic Lupus Erythematosus Patients. *Immunity.* 2010;32:129–140.

[2] Stegle O, Parts L, Durbin R *et al.* A Bayesian framework to account for complex non-genetic factors in gene expression levels greatly increases power in eQTL studies. *PLoS computational biology.* 2010;6:e1000770.

[3] Ongen H, Buil A, Brown AA *et al.* Fast and efficient QTL mapper for thousands of molecular phenotypes. *Bioinformatics (Oxford, England).* 2016;32:1479–85.

[4] Shabalin AA. Matrix eQTL: ultra fast eQTL analysis via large matrix operations. *Bioinformatics (Oxford, England)*. 2012;28:1353–8.

[5] Westra HJ, Peters MJ, Esko T *et al.* Systematic identification of trans eQTLs as putative drivers of known disease associations. *Nature genetics*. 2013;45:1238–43.

[6] Peters JE, Lyons PA, Lee JC *et al.* Insight into Genotype-Phenotype Associations through eQTL Mapping in Multiple Cell Types in Health and Immune-Mediated Disease. *PLoS genetics*. 2016;12:e1005908.

[7] Li Y, Zhang K, Chen H *et al.* A genome-wide association study in Han Chinese identifies a susceptibility locus for primary Sjögren's syndrome at 7q11.23. *Nature genetics*. 2013;45:1361–5.

[8] Abbas aR, Baldwin D, Ma Y *et al.* Immune response in silico (IRIS): immune-specific genes identified from a compendium of microarray expression data. *Genes and immunity*. 2005;6:319–31.

[9] Novershtern N, Subramanian A, Lawton LN *et al.* Densely interconnected transcriptional circuits control cell states in human hematopoiesis. *Cell*. 2011;144:296–309.

[10] Chikina M, Zaslavsky E & Sealfon SC. CellCODE: a robust latent variable approach to differential expression analysis for heterogeneous cell populations. *Bioinformatics*. 2015;31:1584–1591.

[11] Lui TWH, Tsui NBY, Chan LWC *et al.* DECODE: an integrated differential co-expression and differential expression analysis of gene expression data. *BMC bioinformatics*. 2015;16:182.

[12] Makashir SB, Kottyan LC & Weirauch MT. Meta-analysis of differential gene co-expression: application to lupus. *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*. 2015;443–54.

[13] Liberzon A, Birger C, Thorvaldsdóttir H *et al.* The Molecular Signatures Database (MSigDB) hallmark gene set collection. *Cell systems*. 2015;1:417–425.

[14] Merico D, Isserlin R, Stueker O *et al.* Enrichment map: a network-based method for gene-set enrichment visualization and interpretation. *PloS one*. 2010;5:e13984.

[15] Hänzelmann S, Castelo R & Guinney J. GSVA: gene set variation analysis for microarray and RNA-seq data. *BMC bioinformatics*. 2013;14:7.

[16] Imgenberg-Kreuz J, Sandling JK, Almlöf JC *et al.* Genome-wide DNA methylation analysis in multiple tissues in primary Sjögren's syndrome reveals regulatory effects at interferon-induced genes. *Annals of the rheumatic diseases*. 2016;75:2029–2036.

[17] McLean CY, Bristor D, Hiller M *et al.* GREAT improves functional interpretation of cis-regulatory regions. *Nature biotechnology*. 2010;28:495–501.

[18] Langfelder P & Horvath S. WGCNA: an R package for weighted correlation network analysis. *BMC bioinformatics*. 2008;9:559.

[19] Frey BJ & Dueck D. Clustering by passing messages between data points. *Science (New York, N.Y.)*. 2007;315:972–6.

[20] Langfelder P, Luo R, Oldham MC *et al.* Is my network module preserved and reproducible? *PLoS computational biology*. 2011;7:e1001057.