Genome-wide association scan in women with systemic lupus erythematosus identifies susceptibility variants in *ITGAM, PXK, KIAA1542* and other loci

The International Consortium for Systemic Lupus Erythematosus Genetics (SLEGEN)[11], John B Harley[1,2], Marta E Alarcón-Riquelme[3], Lindsey A Criswell[4], Chaim O Jacob[5], Robert P Kimberly[6], Kathy L Moser[1,7], Betty P Tsao[8], Timothy J Vyse[9], Carl D Langefeld[10], Swapan K Nath[1], Joel M Guthridge[1], Beth L Cobb[1], Daniel B Mirel[12], Miranda C Marion[10], Adrienne H Williams[10], Jasmin Divers[10], Wei Wang[10], Summer G Frank[1], Bahram Namjou[1], Stacey B Gabriel[12], Annette T Lee[13], Peter K Gregersen[13], Timothy W Behrens[7,14], Kimberly E Taylor[4], Michelle Fernando[9], Raphael Zidovetzki[15], Patrick M Gaffney[1,7], Jeffrey C Edberg[6], John D Rioux[16], Joshua O Ojwang[1], Judith A James[1], Joan T Merrill[1], Gary S Gilkeson[17], Michael F Seldin[18], Hong Yin[3], Emily C Baechler[7], Quan-Zhen Li[19], Edward K Wakeland[19], Gail R Bruner[1], Kenneth M Kaufman[1,2], and Jennifer A Kelly[1]

[1]Oklahoma Medical Research Foundation, 825 NE 13th Street, Oklahoma City, Oklahoma 73104, USA.

[2]US Department of Veterans Affairs Medical Center, 921 NE 13th Street, Oklahoma City, Oklahoma 73104, USA, and University of Oklahoma Health Sciences Center, 1100 N Lindsey, Oklahoma City, Oklahoma 73104, USA.

[3]University of Uppsala, Olofsgatan St 10B, Uppsala SE-751 05 Sweden.

[4]University of California at San Francisco, 374 Parnassus Ave, San Francisco, California 94143, USA.

[5]University of Southern California, 2011 Zonal Avenue, Los Angeles, California 90033, USA.

[6]University of Alabama at Birmingham, 1900 University Blvd., Birmingham, Alabama 35294, USA.

[7]University of Minnesota, 312 Church Street, Minneapolis, Minnesota 55455, USA.

[8]University of California at Los Angeles,1000 Veterans Avenue, Los Angeles, California 90095, USA.

[9]Imperial College London, Hammersmith Hospital, Du Cane Road, London W12 0NN UK.

[10]Wake Forest University Health Sciences, Medical Center Blvd., Winston-Salem, North Carolina 27157, USA.

[12]Broad Institute, 7 Cambridge Center, Boston, Massachusetts 02142 USA.

[13]Feinstein Institute for Medical Research, North Shore LIJ Health System, 350 Community Drive, Manhasset, New York 11030, USA.

[14]Genentech, Inc.,1 DNA Way, South San Francisco, California 94080, USA.

[15]University of California at Riverside, 900 University Avenue, Riverside, California 92521, USA.

[16]Université de Montréal,5000 rue Belanger, Montréal H1T 1C8, Canada.

[17]Medical University of South Carolina,135 Rutledge Avenue, Charleston, South Carolina 29403, USA.

[18]University of California at Davis,1 Shields Avenue, Davis, California 95616, USA.

[19]University of Texas Southwestern Medical School, 5323 Harry Hines Boulevard, Dallas, Texas 75235, USA.

Abstract

Systemic lupus erythematosus (SLE) is a common systemic autoimmune disease with complex etiology but strong clustering in families (λ_S = ~30). We performed a genome-wide association scan using 317,501 SNPs in 720 women of European ancestry with SLE and in 2,337 controls, and we genotyped consistently associated SNPs in two additional independent sample sets totaling 1,846 affected women and 1,825 controls. Aside from the expected strong association between SLE and the HLA region on chromosome 6p21 and the previously confirmed non-HLA locus *IRF5* on chromosome 7q32, we found evidence of association with replication ($1.1 \times 10^{-7} < P_{overall} < 1.6 \times 10^{-23}$; odds ratio 0.82–1.62)in four regions: 16p11.2 (*ITGAM*), 11p15.5 (*KIAA1542*), 3p14.3 (*PXK*) and 1q25.1 (rs10798269). We also found evidence for association ($P < 1 \times 10^{-5}$) at *FCGR2A*, *PTPN22* and *STAT4*, regions previously associated with SLE and other autoimmune diseases, as well as at 9 other loci ($P < 2 \times 10^{-7}$). Our results show that numerous genes, some with known immune-related functions, predispose to SLE.

SLE (OMIM 152700) is a multisystem, autoimmune inflammatory disease characterized by antinuclear autoantibodies, complement and interferon activation and tissue destruction. The estimated prevalence of SLE is 31 per 100,000 women in populations of European ancestry, which is 50–75% lower than in other populations[1,2]. SLE predominantly affects women (at a 9:1 ratio), particularly during childbearing years, and has strong genetic and environmental components[2–4]. The estimated concordance rate among monozygotic twins (~30%) is ten times that among dizygotic twins (~3%), in accordance with a high sibling relative risk ratio (λ_S = 29)[2,3].

SLE is influenced by genomic variation, and some variants are hypothesized to interact with each other and with environmental factors[3,4]. Replicated linkages with SLE have been reported[5–8] at 1q23–q25, 1q41–q42, 2q35–q37, 4p16–p15, 4q31–q33, 6p21.3, 6p22–p11, 7p22, 16p12–q13, 19q13, 20p13–p12 and 20q12. Replicated associations with SLE have been reported with variants in the HLA region[9], *FCGR3A*[10], *FCGR2A*[11], *PDCD1* (ref. 12), *IRF5* (refs. 13,14) and *PTPN22* (ref. 15). Rare monogenic forms of lupus occur with mutations in *TREX1*, which encodes a DNA exonuclease[16], or with complete deficiencies of the complement components C1q, C2 or C4 (ref. 17). A recent study reported an association between SLE and variants in *STAT4* (ref. 18), which we confirm below.

Existing genome-wide association (GWA) technologies enable agnostic genome-wide searches for variants predisposing to disease. To that end, the Alliance for Lupus Research formed and supported the International Consortium for Systemic Lupus Erythematosus Genetics (SLEGEN). Here, we report results of an SLE case-control GWA study and a large replication experiment, comprising a total of 6,728 women of European ancestry.

The women with SLE evaluated in the GWA scan had typical clinical manifestations. Of the 720 women with SLE, 591 were probands from pedigrees multiplex for SLE. The remaining affected women self-reported having a family history of SLE or other autoimmune disease. The 720 women with SLE and the 2,337 controls, who provided high-quality genotyping data and who seemed to constitute a homogeneous, nonstratified population sample, were together divided into two similarly matched subsets (Sets 1 and 2) that were evaluated separately and jointly. In addition, we studied 8,230 SNPs identified from the initial analyses of Sets 1 and 2 in two additional sets of affected women and controls. Set 3 contained 920 affected women and 819 controls, and Set 4 contained 926 affected women

and 1,006 controls. We report results for the individual sets and for the combined sample. All affected women and controls were of self-reported female gender and European descent.

We required our primary associations to be significant for having the same risk allele at $P <$ 0.05 in each set (1, 2, 3 and 4), and we required joint analysis of all of the data to be significant at $P < 10^{-6}$ (Table 1). The individual sets had good statistical power ($1 - \beta$ 0.99; $\alpha = 10^{-7}$) to detect and replicate associations from common alleles with odds ratios (ORs) > 1.5 or <0.67 and had lower power (~0.25 < $1 - \beta$ < ~0.80) to detect effects with ORs >1.25 or <0.80 (Supplementary Table 1 and Supplementary Fig. 1 online). The joint analysis provided improved power for modest effects (Supplementary Table 1 and Supplementary Fig. 1). The major results of this study (Fig. 1 and Table 1) met these criteria and the statistical and experimental quality control criteria (Supplementary Methods, Supplementary Fig. 2 and Supplementary Table 2 online).

The most significant association was in the HLA region at 6p21.3. In this 7.05-Mb region, 93 SNPs had $P < 10^{-6}$ in the joint analysis and $P < 0.05$ in each sample set (1, 2, 3 and 4) (Fig. 1 and Supplementary Fig. 3 online). This probably represents the long-range linkage disequilibrium (LD) related to the extended HLA-A1-B8-DR3 haplotype, which is not found to such an extent in non-European populations[10]. Our data are consistent with major association effects ($P < 10^{-6}$) from 26.013 Mb to 32.891 Mb (rs7748167–rs7383287) (Fig. 1, Tables 1 and 2 and Supplementary Fig. 3). The most significantly associated SNPs achieve $P < 10^{-50}$ at chromosome 6 positions 31829012 (rs3131379) and 32026839 (rs1270942). As points of reference, the *C4A* and *HLA-DRB1* genes begin at chromosome 6 positions 32.057 Mb and 32.628 Mb, respectively (Fig. 1 and Supplementary Fig. 3).

Logistic regression analysis of the HLA region identified two partially independent effects, which were identical to those identified in the overall logistic regression analysis (Table 2). As genotyping for HLA-DRB1* or other histocompatibility genes was not available for these subjects, this model cannot be considered final.

Outside the HLA region, we found three SNPs associated with SLE in or very near *ITGAM* on 16p11.2 (Table 1). ITGAM (also called CD11b) combines with the β_2 chain (ITGB2) to form a leukocyte integrin (commonly referred to as MAC-1 or complement receptor 3 (CR3)) that is important for adherence of neutrophils and monocytes to stimulated endothelium. ITGAM is also a receptor for the complement component C3 degradation product, iC3b[19]. The associated SNPs in *ITGAM* were in low LD; *ITGAM* haplotypes did not explain additional variation beyond the individual SNP*s*. Another group has independently discovered an association between SLE and *ITGAM*[20] by studying candidate genes in the 16p12–q13 linkage interval. Of the 9,073 samples used in both studies, 33% are shared. The 7,380 independent samples from women of European ancestry used in the two studies produced a combined joint $P = 2.02 \times 10^{-26}$ and OR = 1.65 (95% confidence interval (CI): 1.45–1.88) at rs9888739 and $P = 3.7 \times 10^{-16}$ and OR = 1.42 (95% CI: 1.27–1.61) at rs1143678.

Previous studies have established association between SLE and variants in the *IRF5* and *TNPO3* region on 7q32 (refs. 13,14), which begins at 128.356 Mb. The purported functional SNPs in *IRF5* are not included on the Infinium HumanHap300 arrays and were not evaluated in our study. The three significant markers presented in Table 1 spanned 149 kb and had low r^2 (<0.10) but higher D′ (0.35   D′   1). A logistic regression model (Table 2) incorporated only rs12537284 into models of these three SNPs in *IRF5* and *TNPO3*. To determine whether this association was independent of the previously reported association at this locus, we performed a separate logistic regression analysis on a subset of samples for which genotyping data on both the GWA SNPs (Sets 1 and 2) and previously reported *IRF5*

SNPs (rs752637 and rs729302)[13] were available. Our analysis strongly suggests that the associations with the *IRF5* and *TNPO3* region in this study are driven by a haplotype of rs752637 and rs729302 (data not shown).

We also observed replication with genome-wide significance in the joint analysis at three additional loci. SNP rs4963128 is in *KIAA1542* at 11p15.5 (joint OR = 0.78; $P = 3.0 \times 10^{-10}$), a genomic region homologous to a gene encoding an elongation factor. This SNP is in a region with a reported insertion-deletion polymorphism[21]. In addition, rs4963128 is 23 kb telomeric to *IRF7*, a gene that is important in interferon-α production, and had an $r^2 = 0.94$ with rs709266 in *IRF7*. Second, SNP rs6445975 at position 58345217 in *PXK* showed strong evidence for association ($P = 7.1 \times 10^{-9}$; OR = 1.25). *PXK* at 3p14.3 encodes a Phox homology domain–containing serine-threonine kinase that has five known human splice variants, three of which are expressed in a wide variety of tissues[22]. Third, rs10798269, which lies outside any recognized gene, was also associated with SLE (Table 1; OR = 0.82 and $P = 1.11 \times 10^{-7}$). Notably, this marker is within the 1q23–q25 SLE linkage interval. We provide the haplotype block structures of the major associations (except the HLA region) in Supplementary Figure 4 online.

Using stepwise multiple logistic regression, we modeled the independent contributions of the SNP*s* that (i) were individually significant at $P < 10^{-6}$ in the joint analysis, (ii) had $P < 0.05$ in each set (1, 2, 3 and 4) and (iii) had an OR in the same direction in each set. For this, we modeled 177 HLA SNPs and 33 non-HLA SNPs using $P < 10^{-6}$ as stringent entry and exit criteria. The final model showed that six of the seven consistent association effects presented in Table 1 (OR > 1.2 and $P < 10^{-6}$) made independent contributions to genetic susceptibility for SLE (Table 2). We detected two separate and independent effects in the HLA region and one additional independent effect in each of five of the six remaining genomic regions from Table 1. Considered jointly, these SNP*s* are predictive of SLE (*C* statistic = 0.67). The *C* statistic is the classic receiver operator characteristic (ROC) curve that provides a measure of the variation explained by the variables in the model. This value is comparable to that of other diagnostic tests, such as the prostate-specific antigen for prostate cancer[23].

Alternatively considered, the SNPs listed in Table 2 jointly explained ~15% of the sibling risk ratio of 29. Here, we estimated the penetrance of each SNP from the logistic regression model and applied equations 12 to 14 from ref. 24. This estimate is optimistic, as it was calculated in the same sample used to select the best SNPs.

Of the associated regions in Table 1, only rs10798269 (1q25.1) was not included in the logistic regression model presented in Table 2. When we relaxed entry and exit criteria to $P < 10^{-4}$, this marker also entered and remained in the model. Computing all pairwise interactions via logistic regression models did not provide any evidence of a statistical interaction among the associated SNPs, either within or between genes ($P > 0.01$).

In a separate analysis, when we removed the requirement for replication across all four sample sets, we detected a number of additional possible non-HLA associations (Table 3). This analysis generally identifies SNPs that have relatively common minor allele frequencies (>0.1) and OR values close to 1.2, a reflection of the statistical power of the study (Supplementary Fig. 1). The associations we identified included markers in genomic regions containing *XKR6*, *LYN*, *ATG5*, *ICA1*, *BLK* and *SCUBE1* (Table 3). We identified five associated SNPs in *XKR6* (XK, Kell blood group complex–related protein 6). Two *XKR6* isoforms have been described, although little is currently known about this gene. *LYN* encodes a Src family kinase important in signal transduction. Normal B cell receptor (BCR) stimulation and aggregation activate Lyn to phosphorylate tyrosine residues in

ITAM-containing BCR-associated Igα and Igβ signaling molecules. Some studies have also reported altered LYN levels in individuals with SLE. In addition, increasing or decreasing mouse Lyn produces lupus-like disease. Knockout of the *ATG5* gene (autophagy 5) results in caspase-dependent apoptosis from the FAS and TNF-α ligands. *ICA1* encodes an islet cell antigen (ICA69) expressed in brain, pancreas, salivary and lacrimal glands that acts as a self-antigen in type 1 diabetes and Sjögren's syndrome. B lymphoid tyrosine kinase (BLK) affects functions associated with the pre-BCR. Like *LYN*, *BLK* encodes a member of the Src family and thus may influence cell proliferation and differentiation. SCUBE1 (signal-peptide-CUB (complement proteins C1r/C1s-UEGF-Bmp1-like) domain-EGF-related 1) belongs to the epidermal growth factor superfamily. SCUBE1 expression is rapidly downregulated during endothelial cell activation, for example, by interleukin-1β or TNFα. SCUBE1 is highly expressed in the alpha granules of platelets and is translocated to the cell surface upon activation and aggregation, where it stimulates the release of potent inflammatory, mitogenic and proliferative molecules into the vascular microenvironment.

We also evaluated 20 previously reported associations with SLE or other autoimmune diseases in the SLEGEN GWA data set (Sets 1 and 2) (Table 4). Our data replicated associations in *PTPN22* (ref. 15), *FCGR2A*[10,11] and *STAT4* (ref. 18).

In summary, we present four new regions having genetic associations with SLE in women of European descent: *ITGAM*, *KIAA1542*, *PXK* and rs10798269. In addition, we identify other genomic regions possibly associated with SLE and confirm associations with the HLA region and with *IRF5*, *STAT4*, *FCRG2A* and *PTPN22*. These genetic factors should prove to be important in SLE pathogenesis.

METHODS

Initial SLEGEN GWA sample

All women with SLE satisfied the revised criteria for classification of SLE from the American College of Rheumatology[25]. The SLEGEN GWA sample initially consisted of 730 unrelated women with SLE and 475 controls obtained by SLEGEN members, all self-identified as females of European ancestry (Supplementary Table 3 online). Two female controls were matched by age and self-reported origin to four affected women in block matching. Seven centers contributed samples for the GWA (Supplementary Table 3). When possible, self-reported ancestry was obtained on the basis of grandparental country of origin. We obtained informed consent from participants for all genotyped specimens under the auspices of the appropriate authority at each institution.

SLEGEN+Illumina sample (Sets 1 and 2)

We obtained data from additional female 'out-of-study' controls genotyped on the Infinium HumanHap300 from Illumina's iControlDB (see URLs section below) and added these data to the SLEGEN GWA sample. iControlDB contains genotype information on 3,904 controls of European ancestry, the majority of which (2,300 records) are from the Robert S. Boas Center for Genomics and Human Genetics at the Feinstein Institute for Medical Research. Sixty-three percent (2,444) of the controls are female, and 3,620 have genotyping data available for at least 317,503 SNPs (range: 243,991–561,466 SNPs). A principal component analysis (PCA) for genetic heterogeneity on the SLEGEN GWA sample plus these Illumina controls ('SLEGEN+Illumina') described below identified 112 genetically distinct samples (102 Illumina controls and 10 women with SLE), which we removed from further analysis. The final SLEGEN+Illumina data set consists of 720 affected women and 2,337 controls.

We divided the sample into two independent subsets (Set 1, consisting of 366 women with SLE and 1,164 controls, and Set 2, consisting of 354 women with SLE and 1,173 controls;

Supplementary Table 3) using the classic block randomization approach used in clinical trials. Specifically, the SLEGEN samples matched four affected women to two controls (the 'in-study' controls) based on self-reported geographic origin, age and recruitment center. The 4:2 matching allowed Set 1 and Set 2 to have a 2:1 matching of cases and controls while balancing the covariates both within and between Set 1 and Set 2. Two of the four affected women were randomly assigned to Set 1 and two to Set 2. Similarly, the corresponding two controls were randomly assigned to Set 1 and Set 2. Both Set 1 and Set 2 samples were genotyped on all 317,000 SNPs. 'Out-of-study' Illumina controls were randomized evenly to Set 1 and Set 2. Each set was analyzed separately and in a combined joint analysis. The SNPs identified for further genotyping in the Lupus Large Association Study, described below, were based on the ranking of the results of Set 1 and Set 2 samples before the availability of the 'out-of-study' Illumina controls.

Lupus Large Association Study (LLAS): Sets 3 and 4

The LLAS is a replication cohort from which two sets of samples were drawn. Set 3 consisted of 1,739 samples, 920 independent affected women of European ancestry (577 European Americans and 343 Europeans) and 819 controls of European ancestry (567 European Americans and 252 British 1958 Birth Cohort controls). Set 4 consisted of 1,932 samples: 920 independent affected women of European ancestry (847 European Americans and 79 Europeans) and 1,006 controls of European ancestry (881 European Americans and 125 British Cohort controls). DNA samples for the LLAS were provided as listed in Supplementary Table 3.

Genotyping and laboratory quality control

The SLEGEN GWA DNA samples were stored and processed at the Broad Institute Center for Genotyping and Analysis (CGA). Double-stranded DNA quantity was assessed using PicoGreen (Molecular Probes). The Sequenom platform was used to obtain a 24-SNP marker genotypic fingerprint (including gender confirmation); 23 of the 24 SNPs were also on the Infinium HumanHap300 arrays and served as a cross-platform sample genotype verification. In addition, rs729302 and rs752637 in the fingerprint from *IRF5* were used in the logistic regression discussed in the text.

Genotyping methods followed ref. 26. Approximately 750 ng of genomic DNA was used to genotype each sample on the Illumina Infinium HumanHap300 genotyping BeadChip (Illumina) at the Broad Institute CGA. Samples were processed according to the Illumina Infinium 2 Assay instruction manual. Briefly, each sample was whole-genome amplified, fragmented, precipitated and resuspended in appropriate hybridization buffer. Denatured samples were hybridized on prepared HumanHap300 BeadChips. After hybridization, the BeadChip oligonucleotides were extended by a single labeled base, which was detected by fluorescence imaging with an Illumina Bead Array Reader. Normalized bead intensity data obtained for each sample were loaded into the Illumina BeadStudio 2.0 (and 3.0) software, which converted fluorescence intensities into SNP genotypes. Data from control subjects from the New York Health Project (formerly the New York Cancer Project)[27] who were included in the initial 475 controls had previously been genotyped at the Feinstein Institute; their raw fluorescence intensity files were processed using BeadStudio at the Broad Institute CGA for consistency of genotype recording. HapMap CEU population DNA samples (Coriell) were used as process controls for the Infinium genotyping. We used a call rate of 95% as a minimum threshold for per-sample genotyping completeness. Consequently, there were 1,351 total scans on 1,222 distinct SLEGEN samples in order to improve call rates <95%. Fifty monomorphic SNPs were removed from the analysis.

LLAS samples

The LLAS DNA samples were assembled at the Oklahoma Medical Research Foundation (OMRF), and 250 ng DNA from each subject was genotyped on the Infinium platform using the BeadStation (Illumina) at the Lupus Genetics Studies unit of the OMRF.

The initial SLEGEN GWA samples were genotyped as detailed above, and the top 13,000 SNPs from Set 1 were considered for the LLAS replication study. We removed (i) SNPs that were redundant or presumed surrogates of another, (ii) those that were not predicted to perform well in the Infinium assay and (iii) those that did not pass quality control standards for the data produced (Supplementary Methods). Approximately 250 ng of genomic DNA was used to genotype each sample on the Illumina iSelect multisample genotyping BeadChip. Samples were processed at the OMRF according to the Illumina Infinium II Assay Multi-Sample instruction manual. In brief, the samples were whole-genome amplified and then fragmented, precipitated and stored. After hybridization to the BeadChips, SNPs were extended by single labeled nucleotides, stained with XStain BC2 and read on an Illumina Bead Array Station. Normalized bead intensity data obtained for each sample were loaded into the Illumina BeadStudio 3.1 software, which converted fluorescence intensities into SNP genotypes.

Duplicate samples

Cryptic relationships and duplicates in the GWA samples were identified and removed by computing the proportion of genotypes that matched based on a set of 500 SNPs with minor allele frequency >0.40. Five pairs of duplicates were found using the criteria of matching on 99.9% of the 500 SNPs.

Statistics

After verifying that allele calls were from the same DNA strand, testing for association was completed using the freely available program SNPGWA (see URLs section below). For each SNP, missing data proportions for cases and controls, minor allele frequency and exact tests for departures from Hardy-Weinberg expectations were calculated.

The additive model was used as the primary hypothesis of statistical inference, unless the lack-of-fit test for the additive model was significant ($P < 0.05$). If so, then the minimum P value from the dominant, additive and recessive models was used; for recessive models, at least 30 women homozygous for the minor allele were required. All genetic models were defined relative to the minor allele. We calculated ORs, 95% CI values, P values, sensitivity and specificity and the C statistic as described below.

The 210 strongest associations from the Illumina 317K array (177 from the HLA region and 33 from across the remaining portion of the genome) were entered into a multiple logistic regression model for SLE using the stepwise model building method (that is, forward selection with backward elimination). For the model reported in Table 2, we used equal entry and exit criteria of $P = 1 \times 10^{-6}$.

We computed the area under the ROC curve or the C statistic. In this context, sensitivity is defined as the probability that a woman has the risk polymorphism(s) given that the woman has SLE. Similarly, specificity is defined as the probability that a woman does not have the risk polymorphism(s) given that the woman does not have SLE. The ROC curve plots sensitivity (on the y axis) versus $1 -$ specificity (on the x axis), where the points on the curve correspond to the different thresholds or cut points based on the probabilities of SLE predicted by logistic regression. The C statistic reported here is the area under the ROC curve, which is an estimate of the probability that, for a randomly selected pair of women

(one with SLE and one without SLE), the woman without SLE can be correctly identified, given her SNP genotype data and the results of the logistic regression model. Thus, a C statistic of 0.5 corresponds to a random selection performance, and 1.0 corresponds to perfect selection performance. Note that the C statistic reported is an initial estimate. Specifically, it is upwardly biased because it is based on the data used to identify the SNPs of interest, but it is downwardly biased because the SNPs are in LD only with the functional variation.

Under an additive model, we computed the sibling relative risk for the six SNPs using the methods of ref. 24. From the logistic regression model and the respective genotype and allele frequencies from each SNP, we estimated the penetrance for each SNP. Applying equations 12, 13 and 14 from ref. 24 provides a sibling risk ratio, allowing the computation of the proportion of the sibling risk ratio explained by these SNPs.

To test for any two-locus interactions, we computed a logistic regression model for each pair of SNPs in Tables 1 and 2. The model contained only the SNPs from these sources and their multiplicative interaction. (We did not obtain any evidence for epistasis.) To examine the inflation of the test statistics due to potential sources of bias (for example, population substructure), we compared the χ^2 value from the additive genetic model to its theoretical distribution. In addition, we report the quantile-quantile (Q-Q) plots for Sets 1, 2 and 3 (Supplementary Fig. 2). These plots compare the observed versus expected values of the Z test statistics under the null hypothesis of no association across the genome. The Q-Q plots are reported with and without adjustment for potential admixture using the principal component analysis and genomic control described below. Together, these showed little bias from the null distribution from the test statistic (Supplementary Table 2).

To account for potential confounding substructure or admixture in these samples, we computed PCAs using all SNPs[28,29]. Owing to the computational complexity of computing the covariance matrix using all SNPs, we transposed the subject-by-SNP matrix[28], computed the analysis separately for each chromosome and combined it to obtain the principal component score[29]. The PCA for substructure was computed twice: once for the SLEGEN cases ($n = 730$) and controls ($n = 475$) and once for the SLEGEN cases and all control samples ($n = 2439$). Samples that violated the assumption of sample homogeneity based on the PCA (102 controls and 10 cases) were removed from the analysis. These 102 Illumina controls were of African descent (M.F.S., unpublished data).

The first principal component in this trimmed sample explained 85% of the observed genetic variation. The distributions of the test statistics for the SLEGEN+Illumina sample with and without this adjustment are comparable and only very modestly inflated. Repeating the GWA analysis using the first principal component score as a covariate yielded nearly identical results. Notably, there was no inflation in the Set 3 test statistic mean or variance. Finally, we made a genomic control adjustment to the χ^2 statistics, both with and without the principal component as a covariate (Supplementary Table 2). The SLEGEN GWA design used block randomization to partition cases and controls into two independent subsamples, providing an opportunity for within-study independent replication of strong associations. Before 'out-of-study' controls from Illumina were available, we used the cases from Set 1 and Set 2 and the initially available controls to identify polymorphisms for follow-up in the subsequent replication sample sets. Analyses were computed for the two subsamples and for the entire sample. The overall ranking of the two subsamples was completed using the Euclidean distance (L$_2$-norm) from a point (x,y), where $x = y$ and both are larger than the maximum of the base 10 logarithm of the inverse of the P value in either subsample. These rankings were used to identify SNPs to be typed in the LLAS. The rankings based on the Euclidean distance were highly concordant with the corresponding P

value in the GWA joint analysis (correlation coefficient $r = 0.923$), but the split-sample design allowed independent replication for markers reaching genome-wide significance in both samples and allowed a combined ranking (Supplementary Fig. 5 online). The LLAS samples were genotyped sequentially and denoted as Set 3 and Set 4.

The significant associations remained after adjusting the analysis by the first principal component in logistic regression models and by a genomic control correction factor. After correcting for multiple comparisons using the Q value extension of the false discovery rate (FDR), Sets 1 and 2 yielded 340 ± 71 (95% CI, assuming normality) effects with an FDR 0.05; $73.5\% \pm 8.8\%$ of these SNPs replicated in Set 3 with the same FDR, a very good level of reproducibility.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Footnotes

Acknowledgments

References

1. Danchenko N, Satia JA, Anthony MS. Epidemiology of systemic lupus erythematosus: a comparison of worldwide disease burden. Lupus. 2006; 15:308–318. [PubMed: 16761508]

2. Alarcón-Segovia D, et al. Familial aggregation of systemic lupus erythematosus, rheumatoid arthritis, and other autoimmune diseases in 1,177 lupus patients from the GLADEL cohort. Arthritis Rheum. 2005; 52:1138–1147. [PubMed: 15818688]

3. Deapen D, et al. A revised estimate of twin concordance in systemic lupus erythematosus. Arthritis Rheum. 1992; 35:311–318. [PubMed: 1536669]

4. James JA, et al. An increased prevalence of Epstein-Barr virus infection in young patients suggests a possible etiology for systemic lupus erythematosus. J. Clin. Invest. 1997; 100:3019–3026. [PubMed: 9399948]

5. Tsao BP, et al. Evidence for linkage of a candidate chromosome 1 region to human systemic lupus erythematosus. J. Clin. Invest. 1997; 99:725–731. [PubMed: 9045876]

6. Moser KL, et al. Genome scan of human systemic lupus erythematosus: evidence for linkage on chromosome 1q in African-American pedigrees. Proc. Natl. Acad. Sci. USA. 1998; 95:14869–14874. [PubMed: 9843982]

7. Forabosco P, et al. Meta-analysis of genome-wide linkage studies of systemic lupus erythematosus. Genes Immun. 2006; 7:609–614. [PubMed: 16971955]

8. Lee YH, Nath SK. Systemic lupus erythematosus susceptibility loci defined by genome scan meta-analysis. Hum. Genet. 2005; 118:434–443. [PubMed: 16208513]

9. Graham RR, et al. Specific combinations of HLA-DR2 and DR3 class II haplotypes contribute graded risk for disease susceptibility and autoantibodies in human SLE. Eur. J. Hum. Genet. 2007; 15:823–830. [PubMed: 17406641]

10. Edberg JC, et al. Genetic linkage and association of Fcγ receptor IIIA (CD16A) on chromosome 1q23 with human systemic lupus erythematosus. Arthritis Rheum. 2002; 46:2132–2140. [PubMed: 12209518]

11. Duits AJ, et al. Skewed distribution of IgG Fc receptor IIa (CD32) polymorphism is associated with renal disease in systemic lupus erythematosus patients. Arthritis Rheum. 1995; 38:1832–1836. [PubMed: 8849356]

12. Prokunina L, et al. A regulatory polymorphism within the PD-1 gene is associated with susceptibility to systemic lupus erythematosus. Nat. Genet. 2002; 32:666–669. [PubMed: 12402038]

13. Sigurdsson S, et al. Polymorphisms in the tyrosine kinase 2 and interferon regulatory factor 5 genes are associated with systemic lupus erythematosus. Am. J. Hum. Genet. 2005; 76:528–537. [PubMed: 15657875]

14. Graham RR, et al. A common haplotype of the interferon regulatory factor 5 (IRF5) regulates splicing and expression and is associated with increased risk of systemic lupus erythematosus. Nat. Genet. 2006; 38:550–555. [PubMed: 16642019]

15. Kyogoku C, et al. Genetic Association of the R620W polymorphism of protein tyrosine phosphatase PTPN22 with human SLE. Am. J. Hum. Genet. 2004; 75:504–507. [PubMed: 15273934]

16. Lee-Kirsch MA, et al. Mutations in the gene encoding the 3'-5' DNA exonuclease TREX1 are associated with systemic lupus erythematosus. Nat. Genet. 2007; 39:1065–1067. [PubMed: 17660818]

17. Morgan BP, Walport MJ. Complement deficiency and disease. Immunol. Today. 1991; 12:301–306. [PubMed: 1836729]

18. Remmers EF, et al. STAT4 and the risk of rheumatoid arthritis and systemic lupus erythematosus. N. Engl. J. Med. 2007; 357:977–986. [PubMed: 17804842]

19. Luo BH, Carman CV, Springer TA. Structural basis of integrin regulation and signaling. Annu. Rev. Immunol. 2007; 25:619–647. [PubMed: 17201681]

20. Nath SK, et al. A nonsynonymous functional variant in integrin-α-M (*ITGAM*) is associated with systemic lupus erythematosus (SLE). Nat. Genet. Jan 20.2008 advance online publication, doi: 10.1038/ng.71.

21. Redon R, et al. Global variation in copy number in the human genome. Nature. 2006; 444:444–454. [PubMed: 17122850]

22. Zou X, et al. Expression pattern and subcellular localization of five splice isoforms of human PXK. Int. J. Mol. Med. 2005; 16:701–707. [PubMed: 16142408]

23. Thompson IM, et al. Operating characteristics of prostate-specific antigen in men with an initial PSA level of 3.0 ng/ml or lower. J. Am. Med. Assoc. 2005; 294:66–70.

24. Risch N. Linkage strategies for genetically complex traits. II. The power of affected relative pairs. Am. J. Hum. Genet. 1990; 46:229–241. [PubMed: 2301393]

25. Hochberg MC. Updating the American College of Rheumatology revised criteria for the classification of systemic lupus erythematosus. Arthritis Rheum. 1997; 40:1725. [PubMed: 9324032]

26. Rioux JD, et al. Genome-wide association study identifies new susceptibility loci for Crohn disease and implicates autophagy in disease pathogenesis. Nat. Genet. 2007; 39:596–604. [PubMed: 17435756]

27. Mitchell MK, et al. The New York Cancer Project: rationale, organization, design, and baseline characteristics. J. Urban Health. 2004; 81:301–310. [PubMed: 15136663]

28. Price AL, et al. Principal components analysis corrects for stratification in genome-wide association studies. Nat. Genet. 2006; 38:904–909. [PubMed: 16862161]

29. Narayanaswamy CR, Raghavarao D. Principal component analysis for large dispersion matrices. App. Stat. 1991; 40:309–316.
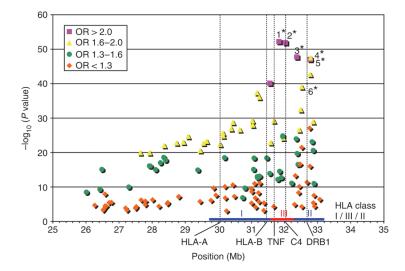
Figure 1.
Combined association results for the extended HLA region (chromosome 6, 26–34 Mb).
SNPs with $P < 0.001$ in the overall joint analysis are represented, color-coded by odds ratio
(OR) strata. SNPs of interest include (i) rs3131379 (position 31.829012 Mb, gene *MSH5*;
OR = 2.36, $P = 1.7 \times 10^{-52}$); (ii) rs1270942 (position 32.026839 Mb, gene *RDBP*; OR 2.35,
$32.369230 = P = 1.3 \times 10^{-51}$); (iii) rs7775397 (position 32.369230 Mb, gene *C6orf10*; OR =
2.28, $P = 8.0 \times 10^{-47}$); (iv) rs9275572 (position 32.786977 Mb, no gene known; OR = 1.69,
$P = 7.0 \times 10^{-48}$); (v) rs1794282 (position 32.774504 Mb, no gene known; OR = 2.26, $P =
2.6 \times 10^{-46}$) and (vi) rs7192 (position 32.519624 Mb, gene *HLA-DRA*; OR = 1.61, $P = 6.2
\times 10^{-40}$).

Table 1

GWA markers associated with SLE

Gene	SNP	Chromosome	Position (Mb)	MAF (affected women)	MAF (controls)	Minor alleles	Set 1 P	Set 2 P	Set 3 P	Set 4 P	Overall joint P	OR	95% CI
HLA region	rs3131379	6p21.33	31.829	0.197	0.103	A	8.02E-13	9.89E-14	1.06E-16	5.63E-12	1.71E-52	2.36	2.11—2.64
HLA region	rs1270942	6p21.32	32.027	0.197	0.104	G	6.22E-12	1.11E-14	8.11E-17	1.72E-11	1.27E-51	2.35	2.10—2.63
ITGAM	rs9888739	16p11.2	31.221	0.189	0.127	T	1.07E-04	2.16E-07	1.84E-06	3.00E-10	1.61E-23	1.62	1.47—1.78
ITGAM	rs1143678	16p11.2	31.251	0.214	0.168	T	3.21E-03	4.78E-03	9.33E-06	2.12E-08	8.50E-14	1.40	1.28—1.53
ITGAM	rs4548893	16p11.2	31.272	0.253	0.208	A	0.01	4.78E-03	1.84E-04	3.43E-08	2.36E-12	1.34	1.24—1.46
IRF5/TNPO3	rs729302	7q32	128.356	0.261	0.315	C	4.28E-03	3.37E-03	1.07E-06	4.75E-04	2.00E-10	0.78	0.72—0.84
IRF5/TNPO3	rs10279821	7q32	128.471	0.267	0.325	T	0.05	9.69E-06	4.42E-06	5.16E-03	6.50E-09	0.80	0.74—0.86
IRF5/TNPO3	rs12537284	7q32	128.505	0.19	0.132	A	1.36E-05	2.06E-07	1.51E-04	1.08E-06	3.61E-19	1.54	1.40—1.70
KIAA1542	rs4963128	11p15.5	0.58	0.288	0.337	T	2.48E-03	3.72E-03	2.42E-03	1.38E-05	3.00E-10	0.78	0.73—0.85
PXK	rs6445975	3p14.3	58.345	0.323	0.276	C	1.80E-02	3.97E-02	1.19E-02	1.84E-05	7.10E-09	1.25	1.16—1.35
-	rs10798269	1q25.1	171.576	0.304	0.356	T	0.04	1.66E-03	1.80E-03	0.023	1.11E-07	0.82	0.76—0.88

SNPs are listed that had P < 0.05 in each of the four sets and that exceeded $P < 10^{-6}$ in the overall joint analysis. For the HLA, only the two most strongly associated SNPs are presented (Fig. 1 and Supplementary Fig. 3). The additive model is presented unless the test for lack of fit to an additive model was significant (P < 0.05), as for rs729302 in Set 1 (recessive). A secondary haplotype analysis did not improve the probabilities presented above for individual SNPs. MAF, minor allele frequency. Odds ratio (OR) is shown with 95% confidence interval (CI) boundaries. Positions are from build 35 throughout.

Table 2

Stepwise logistic regression model of the independent genetic contributions of the associated markers to SLE risk

Gene	SNP	Chromosome	OR	95% CI	*P*
PXK	rs6445975	3p14.3	1.27	1.15–1.39	9.2E-07
HLA region	rs3131379	6p21.33	1.82	1.58–2.09	4.5E-17
HLA region	rs9275572	6p21.32	1.40	1.27–1.54	2.8E-12
IRF5/TNPO3	rs12537284	7q32.1	1.61	1.42–1.81	1.7E-14
KIAA1542	rs4963128	11p15.5	0.78	0.71–0.85	1.3E-07
ITGAM	rs9888739	16p11.2	1.70	1.51–1.92	1.9E-18
C statistic					0.67

Data used for logistic regression are from the 210 SNP$_S$ available from Sets 1–4. Entry or removal from the model at each step required $P < 10^{-6}$. Abbreviations are the same as in Table 1.

Table 3

Overall joint probabilities from non-HLA (6p21) genomic regions suggestive of genetic association

Gene	SNP	Chr	Position (Mb)	MAF (affected women)	MAF (controls)	Minor allele	Set 1 P	Set 2 P	Set 3 P	Set 4 P	Overall joint P	OR	95% CI
NMNAT2	rs2022013	1q25.3	181.620476	0.3797	0.4197	G	0.27761	0.00061	0.04818	0.00002	1.08E-07	0.85	0.79–0.9
-	rs2431697	5q33.3	159.812556	0.3848	0.4342	C	0.00180	0.05430	0.00093	0.01260	1.00E-10	0.82	0.77–0.87
-	rs6568431	6q21	106.695499	0.4237	0.3812	A	0.00066	0.01535	0.00797	0.26498	1.74E-08	1.19	1.12–1.27
ATG5	rs573775	6q21	106.871559	0.3178	0.281	T	0.03760	0.00005	0.00427	0.95153	1.36E-07	1.19	1.12–1.27
ICA1	rs10156091	7p21.3	8.153619	0.1198	0.0969	A	0.03768	0.00003	0.03378	0.10622	1.90E-07	1.32	1.19–1.47
TNPO3	rs10239340	7q32.1	128.455746	0.324	0.3842	T	0.04340	3.73E-07	7.42E-06	0.10424	6.98E-16	0.77	0.73–0.82
-	rs6601327	8p23.1	9.432942	0.4039	0.3649	C	0.00059	0.06181	0.38560	0.01097	1.99E-07	1.18	1.11–1.25
XKR6	rs6985109	8p23.1	10.798995	0.5138	0.462	G	0.00123	0.09025	0.00082	0.00430	2.51E-11	1.23	1.16–1.3
XKR6	rs4240671	8p23.1	10.805158	0.4499	0.4947	A	0.34769	0.12244	0.00552	0.00347	6.60E-09	0.75	0.68–0.83
XKR6	rs11783247	8p23.1	10.826285	0.5099	0.4627	C	0.00469	0.10787	0.00150	0.02095	8.00E-10	1.21	1.14–1.28
XKR6	rs6984496	8p23.1	10.833503	0.5078	0.4588	C	0.00164	0.13155	0.00066	0.02292	2.00E-10	1.22	1.14–1.29
C8orf12	rs7836059	8p23.1	11.309574	0.439	0.4871	T	0.00606	0.08327	0.01088	0.00647	4.00E-10	0.82	0.78–0.88
BLK	rs2248932	8p23.1	11.429059	0.3928	0.3475	T	0.00751	0.45841	0.06147	0.00001	7.00E-10	1.22	1.14–1.3
-	rs10903340	8p21.1	11.487996	0.4593	0.4193	C	0.00346	0.31025	0.07304	0.02869	1.46E-07	1.18	1.11–1.25
LYN	rs7829816	8q12	57.011940	0.1844	0.2159	C	0.52653	0.00187	0.00016	0.95150	5.40E-09	0.77	0.70–0.84
LYN	rs2667978	8q12	57.060505	0.19	0.2244	C	0.00021	0.02071	0.00701	0.51240	5.10E-08	0.81	0.76–0.88
UBE2L3	rs5754217	22q11.21	20.269675	0.2287	0.1953	A	0.01589	0.00187	0.20201	0.01842	7.53E-08	1.22	1.14–1.32
SCUBE1	rs2071725	22q13.2	41.939704	0.1176	0.1453	A	0.00255	0.01762	0.00015	0.84821	1.21E-07	0.78	0.72–0.86

Listed are SNPs that exceed $P < 2 \times 10^{-7}$ in the overall joint analysis but not $P < 0.05$ in each individual set, in contrast to Table 1. The additive model is presented unless the test for lack of fit to an additive model was significant ($P < 0.05$), as for rs7829816 and rs1015609 (dominant). Chr, chromosome; MAF, minor allele frequency. Odds ratio (OR) is shown with 95% confidence interval (CI) boundaries. Positions are from build 35 throughout.

Table 4

Analysis of candidate genes from previous studies of SLE and other autoimmune disorders

Gene	Phenotype	Chr	Published SNP	Minor allele	MAF (affected women)	MAF (controls)	GWA P value	OR	95% CI
IL23R	CD	1p31.3	rs11209026	T	0.0671	0.0665	0.93	1.01	0.72–1.41
PTPN22	SLE, RA, T1D, AITD	1p13	rs2476601	A	0.1271	0.0862	5.2×10^{-6}	1.53	1.27–1.84
FCGR2A	SLE	1q23	rs1801274	T	0.4319	0.5067	6.78×10^{-7}	0.74	0.65–0.83
IFIH1	T1D	2q24	rs1990760	C	0.3706	0.4067	0.01	0.86	0.76–0.97
			rs3788964	G	0.1411	0.1544	0.22	0.90	0.76–1.07
STAT4	RA, SLE	2q32	rs7574865	T	0.3159	0.2346	2.8×10^{-9}	1.50	1.24–1.82
			rs7601754	C	0.1340	0.1871	3.7×10^{-6}	0.67	0.57–0.80
CAPSL	T1D	5p13	rs1445898	T	0.4451	0.4418	0.83	0.96	0.81–1.14
IL7R	MS, T1D	5p13	rs6897932	A	0.2556	0.2488	0.6	1.04	0.90–1.19
			rs1494558	T	0.3340	0.3387	0.74	0.98	0.82–1.17
			rs1494555	C	0.3299	0.3330	0.83	0.99	0.83–1.18
			rs3194051	C	0.2479	0.2738	0.05	0.88	0.76–1.00
SLC22A4	RA	5q31	rs2073838	A	0.0801	0.0773	0.73	1.04	0.76–1.41
TRAF1-C5	RA	9q33	rs7035682	A	0.0830	0.0841	0.89	0.99	0.73–1.34
			rs10985112	A	0.0751	0.0748	0.97	1.00	0.73–1.38
			rs7026551	G	0.1957	0.1737	0.002	1.90	1.25–2.88
			rs2269066	A	0.1035	0.0887	0.09	1.19	0.97–1.45
IL2RA	T1D, MS	10p15	rs7090530	G	0.4366	0.4124	0.10	1.10	0.98–1.25
			rs12251307	A	0.1292	0.1239	0.6	1.05	0.88–1.25
			rs7072793	C	0.3806	0.4165	0.01	0.86	0.76–0.97
			rs4147359	T	0.3090	0.3403	0.03	0.87	0.77–0.99
SH2B3	T1D	12q24	rs3184504	C	0.4588	0.4922	0.03	0.88	0.78–0.99
PTPN2	T1D	18p11	rs1893217	C	0.1625	0.1577	0.66	1.04	0.83–1.30
			rs478582	G	0.4638	0.4390	0.1	1.10	0.98–1.24
TYK2	SLE	19p13	rs280500	C	0.1557	0.1572	0.9	0.99	0.78–1.24
			rs6511696	A	0.0817	0.0788	0.72	1.03	0.76–1.40
			rs280519	A	0.5258	0.4915	0.03	1.14	1.01–1.28
			rs12720356	G	0.0744	0.0881	0.1	0.83	0.66–1.04

Shown here are results from our GWA data from the joint analysis of Sets 1 and 2 for previously described autoimmune-associated SNP$_S$. We also observed significant associations with rs12580100 ($P = 0.0009$), which is 35 kb from *ERBB3* (T1D) and rs10800309 ($P = 0.0007$), which is 3.3 kb upstream from *FCGR2A* (SLE). Genes not evaluated include *FCGR3A* (SLE), *PDCD1* (SLE), *CARD15* (CD), *CTLA4/ICOS* (SLE, T1D), *CRP* (SLE) and *FCRL3* (RA, SLE). The additive model is presented except for rs7026551 where the lack of fit to an additive model was significant ($P < 0.05$) and a recessive model is preferred. Abbreviations for phenotypes are as follows: AITD, autoimmune thyroid disease; CD, Crohn's disease; MS, multiple sclerosis; RA, rheumatoid arthritis; SLE, systemic lupus erythematosus; T1D, type 1 diabetes. Other abbreviations are as listed in Table 1.