








Global discovery of lupus genetic risk variant allelic enhancer activity

Xiaoming Lu ^{1,14}, Xiaoting Chen ^{1,14}, Carmy Forney¹, Omer Donmez¹, Daniel Miller¹, Sreeja Parameswaran¹, Ted Hong^{1,2,10}, Yongbo Huang¹, Mario Pujato^{3,11}, Tareian Cazares⁴, Emily R. Miraldi ^{3,4,5}, John P. Ray ^{6,12}, Carl G. de Boer ^{6,13}, John B. Harley^{1,4,5,7,8}, Matthew T. Weirauch ^{1,3,5,8,9,15}✉ & Leah C. Kottyan ^{1,4,5,9,15}✉

Genome-wide association studies of Systemic Lupus Erythematosus (SLE) nominate 3073 genetic variants at 91 risk loci. To systematically screen these variants for allelic transcriptional enhancer activity, we construct a massively parallel reporter assay (MPRA) library comprising 12,396 DNA oligonucleotides containing the genomic context around every allele of each SLE variant. Transfection into the Epstein-Barr virus-transformed B cell line GM12878 reveals 482 variants with enhancer activity, with 51 variants showing genotype-dependent (allelic) enhancer activity at 27 risk loci. Comparison of MPRA results in GM12878 and Jurkat T cell lines highlights shared and unique allelic transcriptional regulatory mechanisms at SLE risk loci. In-depth analysis of allelic transcription factor (TF) binding at and around allelic variants identifies one class of TFs whose DNA-binding motif tends to be directly altered by the risk variant and a second class of TFs that bind allelically without direct alteration of their motif by the variant. Collectively, our approach provides a blueprint for the discovery of allelic gene regulation at risk loci for any disease and offers insight into the transcriptional regulatory mechanisms underlying SLE.

¹Center for Autoimmune Genomics and Etiology, Cincinnati Children's Hospital Medical Center, Cincinnati, OH, USA. ²Department of Pharmacology and Systems Physiology, University of Cincinnati, College of Medicine, Cincinnati, OH, USA. ³Division of Biomedical Informatics, Cincinnati Children's Hospital Medical Center, Cincinnati, OH, USA. ⁴Division of Immunobiology, Cincinnati Children's Hospital Medical Center, Cincinnati, OH, USA. ⁵Department of Pediatrics, University of Cincinnati College of Medicine, Cincinnati, OH, USA. ⁶Broad Institute of Massachusetts Institute of Technology (MIT) and Harvard University, Cambridge, MA, USA. ⁷US Department of Veterans Affairs Medical Center, Cincinnati, OH, USA. ⁸Division of Developmental Biology, Cincinnati Children's Hospital Medical Center, Cincinnati, OH, USA. ⁹Division of Allergy and Immunology, Cincinnati Children's Hospital Medical Center, Cincinnati, OH, USA. ¹⁰Present address: Translational Medicine, R&D Oncology, AstraZeneca, Boston, MD, USA. ¹¹Present address: Production Informatics, Oncology, AstraZeneca, Gaithersburg, MD, USA. ¹²Present address: Systems Immunology, Benaroya Research Institute at Virginia Mason, Seattle, Washington, USA. ¹³Present address: School of Biomedical Engineering, University of British Columbia, Vancouver, BC, Canada. ¹⁴These authors contributed equally: Xiaoming Lu, Xiaoting Chen. ¹⁵These authors jointly supervised this work: Matthew T. Weirauch, Leah C. Kottyan. ✉email: Matthew.Weirauch@cchmc.org; Leah.Kottyan@cchmc.org

Systemic lupus erythematosus (SLE) is an autoimmune disease that can affect multiple organs, leading to debilitating inflammation and mortality¹. Up to 150 cases are found per 100,000 individuals, and the limited treatment options contribute to considerable economic and social burden^{1,2}. Epidemiological studies have established a role for both genetic and environmental factors in the development of SLE². SLE has a relatively high heritability³. The vast majority of patients do not have a single disease-causing mutation (such as mutations in complement protein 1q); instead, genetic risk is accumulated additively through many genetic risk loci with modest effect sizes⁴.

Genome-wide association studies (GWASs) have identified 91 genetic risk loci that increase disease risk of SLE in a largely additive fashion⁴. Each SLE-risk locus is a segment of the genome containing a polymorphic “tag” variant (i.e., the variant with the most significant GWAS *p*-value) and the genetic variants in linkage disequilibrium with the tag variant. The majority (68%) of the established SLE-risk loci do not contain a disease-associated coding variant that changes amino acid usage⁵. Instead, variants at these loci are found in non-coding regions of the genome such as introns, promoters, enhancers, and other intergenic areas. Enrichment of these variants in enhancers and at transcription factor (TF) binding sites^{6,7} implies that transcriptional perturbation may be a key to the development of SLE⁸. However, given the large number of candidate variants identified by GWASs, identification of the particular causal variant(s) remains challenging.

SLE is a complex disease that involves multiple cell types². Previous systematic studies demonstrate that SLE-risk loci are enriched for B cell-specific genes⁹ and regulatory regions¹⁰. Established biological mechanisms further highlight a key role for B cells in SLE—as the autoantibody-secreting cell type, B cells are critical to the pathoetiology of SLE, a disease characterized by autoantibody production¹¹. B cells also present self-antigens to T cells in the development of an autoantigen-focused (i.e., “self”) inflammatory response¹². Meanwhile, Epstein–Barr virus (EBV)-infected B cells have been implicated in SLE, with patients having a greater number of EBV-infected B cells and a higher viral load than people without SLE^{13,14}. In addition, EBV infection is significantly more prevalent in SLE cases than controls^{15,16}, and EBV-encoded EBNA2 interactions with the human genome are concentrated at SLE-risk loci in EBV-transformed B cell lines¹⁰. In vitro, EBV infection can transform B cells into a lymphoblastoid cell line (LCL)¹⁷. We have recently shown that histone mark and human and viral protein chromatin immunoprecipitation followed by sequencing (ChIP-seq) datasets from EBV-transformed B cell lines are highly and specifically enriched at SLE-risk loci relative to non-EBV-transformed B cell lines^{9,10}. Given the above evidence, we chose the EBV-transformed B cell line GM12878 to study the effects of SLE-risk variants at SLE-risk loci.

In this work, we design and apply a massively parallel reporter assay (MPRA)^{18–27} to systematically identify the SLE genetic risk variants that contribute to transcriptional dysregulation in the EBV-transformed B cell line GM12878 (Fig. 1 and Supplementary Fig. 1). MPRA extends standard reporter assays, replacing low-throughput luciferase with high-throughput mRNA expression detection. We use MPRA to simultaneously screen the full set of genome-wide significant SLE-associated genetic variants for effects on gene regulation. Using this experimental approach, we nominate 51 putative causal variants that result in genotype-dependent (allelic) transcriptional regulation. Comparison of MPRA results between GM12878 and the Jurkat T cell line reveals shared and cell-type-specific allelic behavior. Integration of these data with TF binding site predictions and functional genomics data reveals two distinct mechanisms whereby TFs bind

risk variants in an allelic manner—directly impacted by a given variant (i.e., the variant directly alters the TF’s DNA-binding site) or indirectly impacted by the variant (i.e., the variant alters the DNA binding of the TF’s physical interaction partner or modulates chromatin accessibility). Collectively, these results provide an important resource for understanding SLE disease risk mechanisms and reveal an important role for groups of TFs in the mediation of allelic enhancer activity at plausibly causal SLE-risk variants in EBV-transformed B cells.

Results

MPRA library design and quality control. We first collected all SLE-associated risk loci reaching genome-wide association significance ($p < 5 \times 10^{-8}$) published through March 2018 (Supplementary Data 1). Studies of all ancestral groups were included, and independent risk loci were defined as loci with lead (tag) variants at $r^2 < 0.2$. For each of these 91 risk loci, we performed linkage disequilibrium (LD) expansion ($r^2 > 0.8$) in each ancestry of the initial genetic association(s), to include all possible disease-relevant variants (Supplementary Data 2). In total, this procedure identified 3073 genetic variants. All established alleles of these variants were included, with 36 variants having three or more alleles. We also included 20 additional genetic variants from a previously published study¹⁹ as positive and negative controls to assess the library’s performance (Supplementary Data 3).

For each variant, we generated a pair of 170 base pair (bp) DNA oligonucleotides (subsequently referred to as “oligos”) for each allele, with the variant located in the center and identical flanking genomic sequence across the alleles (Supplementary Data 4). A total of 12,478 oligos (3093 variants with 6239 alleles) were synthesized. For barcoding, a random 20mers were added to each oligo through PCR. Each unique barcode was matched with perfectly synthesized oligos. The number of unique barcodes per oligo had an approximately normal distribution with a median of 729 barcodes per oligo (Supplementary Fig. 2a and Supplementary Data 5). Only oligos with at least 30 unique barcodes were used for downstream analyses. A fragment containing a minimal promoter and an *eGFP* gene was inserted between the oligo and barcode to generate the MPRA transfection library. We note that the use of a minimal promoter allows us to effectively measure the ability of alleles to enhance, but not reduce, transcriptional activity. Three aliquots of the library were independently transfected into the EBV-transformed B cell line GM12878. We then used nucleic acid capture to enrich for *eGFP* mRNA and sequenced the barcode region. The normalized barcode ratio between the *eGFP* mRNA and the plasmid DNA was used to quantify the amount of enhancer activity driven by each oligo (Supplementary Note 1 and Supplementary Fig. 5f, g). This mRNA to DNA ratio measures the enhancing effect of an allele on *eGFP* expression under the control of a minimal promoter (Fig. 1 and Supplementary Fig. 1). We observed strong correlation of enhancer activity between experimental replicates (mean pairwise Pearson correlation of 0.99) (Supplementary Fig. 2b–d). Likewise, calibration variants showed high accuracy, with 17 of the 20 variants matching the results of a previous study¹⁹ (87.5% sensitivity and 75% specificity), collectively demonstrating a robust experimental system (Supplementary Data 3).

Hundreds of SLE-risk variants are located in genomic regions with enhancer activity in EBV-transformed B cells. Using the SLE MPRA library, we next identified genetic variants capable of driving enhancer activity in the EBV-transformed B cell line GM12878. An SLE-risk variant was considered a candidate for enhancer activity if an oligo corresponding to any allele had

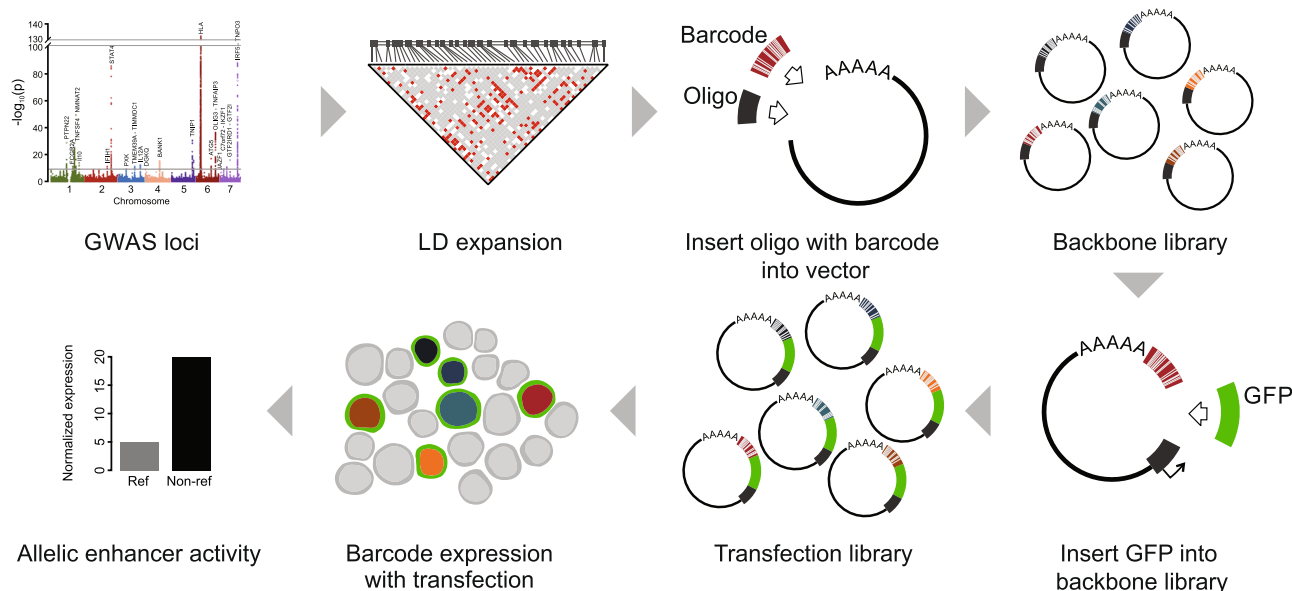


Fig. 1 Massively parallel reporter assay workflow. Schematic of study design. Representative Manhattan plot of SLE-associated risk loci reproduced from ref. 78.

significantly increased transcriptional regulatory activity compared to controls (see “Methods”). Not all statistically significant changes in transcriptional activity are necessarily biologically relevant—a highly consistent, but slight change in expression levels is statistically, but not biologically, meaningful. We therefore considered an oligo to have enhancer activity only when (1) the oligo had statistically significant enhancer activity ($p_{adj} < 0.05$) and (2) we observed at least a 50% increase in transcriptional activity compared to the corresponding barcode counts in the plasmid control. Based on these criteria, 16% of SLE-risk variants (482 variants, 853 alleles) demonstrated enhancer activity, henceforth referred to as “enhancer variants” (enVars) and “enhancer alleles” (enAlleles), respectively (Fig. 2a and Supplementary Data 6).

We next explored the potential effects of enVars on gene expression. We connected each enVar to one or more genes using an approach that takes into account chromatin looping interactions, expression quantitative trait loci (eQTLs), and gene proximity (Supplementary Data 2 and 7) (see “Methods”). This approach identified 1006 genes in total, which are enriched for expected SLE-related processes such as the interferon pathway, the antigen processing and presentation pathway, and cytokine-related pathways (Supplementary Fig. 3 and Supplementary Data 8), providing functional support for the enVars we identified.

Next, we searched for functional genomic features enriched within enVars relative to non-enVars using the RELI algorithm¹⁰. In brief, RELI estimates the significance of the intersection between an input set of genomic regions (e.g., enVars) and each member of a collection of functional genomics datasets (e.g., ChIP-seq for a particular histone mark or TF). For this analysis, we identified, curated, and systematically processed the 576 GM12878 ChIP-seq datasets available in the NCBI Gene Expression Omnibus (GEO) database (see “Methods”). Using RELI, we observed significant enrichment for overlap between enVars and multiple histone modification marks, including H3K4me3 (5.8-fold, $p_{corrected} < 10^{-21}$) and H3K27ac (2.0-fold, $p_{corrected} < 10^{-13}$) (Fig. 2b and Supplementary Data 9). As expected, we did not identify enrichment for repressive marks such as H3K9me3 or H3K27me3²⁸ (Fig. 2b and Supplementary Data 9). Altogether, the genomic features present within enVars

confirm that many SLE genetic risk loci likely alter transcriptional regulation in EBV-transformed B cells.

We next asked if the genomic binding sites of particular TFs were enriched within our enVars using RELI and the TF ChIP-seq datasets from GM12878. As expected, the enVars are highly enriched for ChIP-seq signal of TFs involved in regulation of the immune response, relative to variants lacking enhancer activity (Fig. 2c and Supplementary Data 9). In particular, we found significant enrichment for all members of the NFκB TF family: REL/C-Rel (6.4-fold, $p_{corrected} < 10^{-26}$), NFKB1/p50 (3.0-fold, $p_{corrected} < 10^{-18}$), RELA/p65 (3.1-fold, $p_{corrected} < 10^{-16}$), RELB (2.7-fold, $p_{corrected} < 10^{-10}$), and NFKB2/p52 (2.2-fold, $p_{corrected} < 10^{-7}$). These results are consistent with our previous findings that altered binding of NFκB TFs is likely an important mechanism conferring SLE risk¹⁰. We also found significant enrichment for other TFs that have been previously implicated in SLE pathogenesis, such as PAX5²⁹, MED1³⁰, IKZF1³¹, ELF1³², and the EBV-encoded EBNA2 transactivator¹⁰ (Fig. 2c and Supplementary Data 9). As a complementary approach, we next assessed enrichment for TF binding site motifs in the enAllele DNA sequences using HOMER³³ and motifs contained in the Cis-BP database³⁴ (see “Methods”). This analysis also revealed enrichment of multiple TF families with known roles in SLE, including ETS, NFκB, and IRF³ (Fig. 2d and Supplementary Data 10). Many of these same TFs also have enriched ChIP-seq peaks at SLE-risk loci¹⁰. Collectively, these results indicate that particular TFs tend to not only concentrate at SLE-risk loci¹⁰, but also concentrate at alleles capable of driving gene expression in EBV-transformed B cells.

MPRA identifies 51 SLE-risk variants with allelic enhancer activity in EBV-transformed B cells. We next used our MPRA library to identify SLE genetic risk variants that drive allele-dependent (allelic) enhancer activity. Allelic activity was assessed for each enVar by comparing enhancer activity between each pair of alleles. We considered a SLE variant allelic if (1) at least one of its alleles is an enAllele; (2) we observed significant genotype-dependent activity using Student’s *t*-test^{19,35} (Supplementary Note 2 and Supplementary Fig. 6); and (3) the oligos had more than a 25% change between any pair of alleles. Using these

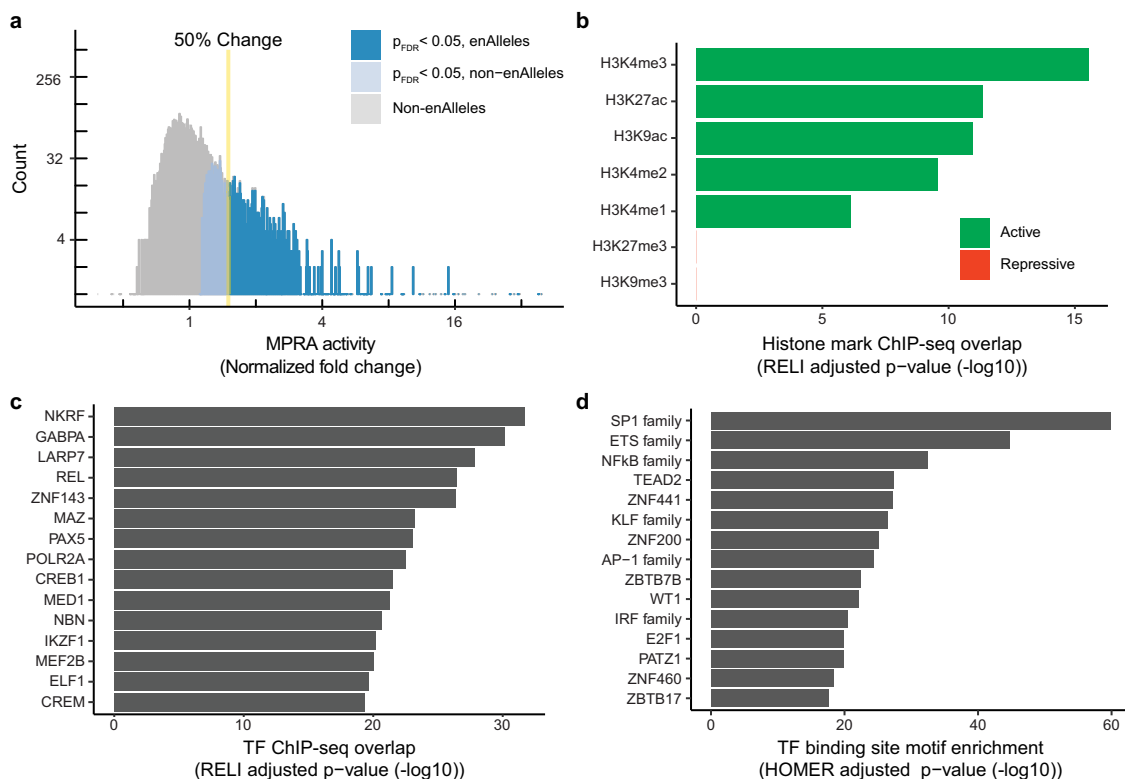


Fig. 2 Regulatory activity of enhancer variants (enVars). **a** Distribution of MPRA regulatory activity. The normalized fold change of MPRA activity relative to plasmid control (X -axis) was calculated using DESeq2 ($n = 3$ biological replicates). Enhancer alleles (enAlleles) (blue) were identified as those alleles with significant activity relative to control ($p_{\text{adj}} < 0.05$) and at least a 50% increase in activity (see “Methods”). The p -values were generated by two-sided Wald tests with Benjamini-Hochberg multiple testing correction. Full results are provided in Supplementary Data 6. **b** Enrichment of histone marks in GM12878 cells at enVars compared to non-enVars. p -values were estimated by one-sided z-test with Bonferroni multiple testing correction using RELI (see “Methods”). Full results are provided in Supplementary Data 9. **c** Enrichment of regulatory protein and transcription factor (TF) binding at enVars compared to non-enVars. p -values were estimated by one-sided z-test with Bonferroni multiple testing correction using RELI (see “Methods”). The top 15 TFs (based on RELI p -values) that overlap at least 10% of enVars are shown. Full results are provided in Supplementary Data 9. **d** TF binding site motif enrichment for enVars compared to non-enVars. p -values were estimated by one-sided hypergeometric test with Benjamini-Hochberg multiple testing correction by HOMER using the full oligo sequences of enVars and non-enVars (see “Methods”). The top 15 enriched TF motif families are shown. Full results are provided in Supplementary Data 10.

criteria, we identified 51 SLE-risk variants (11% of enVars, 1.7% of all SLE-risk variants) as allelic enVars in GM12878 (Fig. 3a and Supplementary Data 11). For 31 of these 51 allelic enVars, the risk allele decreased enhancer activity relative to the non-risk allele, which is statistically indistinguishable from the 20 variants with increased risk allele activity ($p = 0.1$). Three of the allelic enVars can also alter the amino acid sequence of proteins—rs1059702 (IRAK1), rs1804182 (PLAT), and rs3027878 (HCFC1), consistent with previous studies identifying dual-use codons in the human genome³⁶. Collectively, these 51 variants represent causal variant candidates for 27 SLE-risk loci (30% of all tested loci) (Supplementary Data 12). For these 27 risk loci, our approach reduced the number of potential causal variants with allelic activity in GM12878 from an average of 84 variants to an average of two variants per risk locus (Fig. 3b). For example, at 17q12 (marked by rs8079075), we reduce the candidate causal variant set from 249 to one, with the rs112569955 “G” risk allele showing a 36% increase in enhancer activity compared to the “A” non-risk allele.

Particular TFs have altered binding at SLE loci with allelic enhancer activity. To identify candidate regulatory proteins that might participate in allelic SLE mechanisms, we next used RELI to identify GM12878 ChIP-seq datasets that significantly overlap allelic enVars (Supplementary Data 13). Many of the top

results are consistent with our previous study¹⁰, including the enriched presence of general enhancer features such as the H3K27ac histone mark (17 of 51 allelic enVars, 13.6-fold enriched, $p_{\text{corrected}} < 10^{-38}$), mediator complex subunit MED1 (17 of 51 allelic enVars, 13.0-fold enriched, $p_{\text{corrected}} < 10^{-34}$), and the histone acetyltransferase p300 (16 of 51 allelic enVars, 12.4-fold enriched, $p_{\text{corrected}} < 10^{-32}$), along with particular regulatory proteins that participate in “EBV super enhancers”³⁷ and play key roles in B cells such as ATF7 (15 of 51 allelic enVars, 11.3-fold enriched, $p_{\text{corrected}} < 10^{-26}$), Ikaros/IKZF1 (19 of 51 allelic enVars, 9.7-fold enriched, $p_{\text{corrected}} < 10^{-25}$), and the NF κ B subunit RELA (13 of 51 allelic enVars, 12.4-fold enriched, $p_{\text{corrected}} < 10^{-24}$). Also consistent with our previous study¹⁰, we observe strong enrichment for the EBV-encoded EBNA2 protein (7 of 51 allelic enVars, 17.7-fold enriched, $p_{\text{corrected}} < 10^{-19}$). Collectively, these data reveal particular regulatory proteins that might participate in the mechanisms contributing to SLE at multiple risk loci by driving allelic enhancer activity.

We next used the MARIO pipeline¹⁰ to search for allelic binding events (i.e., allelic imbalance between sequencing read counts) at SLE variants within 1058 LCL ChIP-seq datasets (576 from GM12878). By necessity, this approach is limited to the 47 allelic enVars that are heterozygous in at least one of these cell lines. In total, this procedure identified 11 variants with strong allelic imbalance (MARIO ARS value > 0.4) in at least

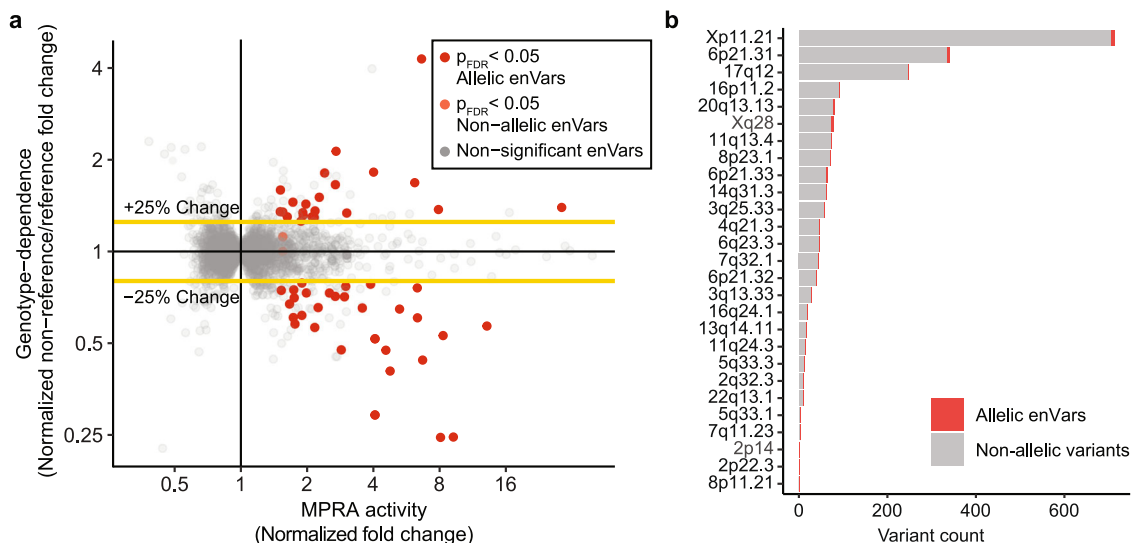


Fig. 3 Regulatory activity of allelic enhancer variants (allelic enVars). **a** Identification of allelic enVars. Genotype dependence (Y-axis) is defined as the normalized fold change of MPRA activity between the non-reference and reference alleles ($n = 3$ biological replicates, see “Methods”). MPRA activity (X-axis) is presented as the maximum normalized fold change of MPRA activity for any allele of the variant. Allelic enVars (red) were defined as variants with a significant difference in MPRA activity ($p_{\text{adj}} < 0.05$) between any pair of alleles and at least a 25% change in activity difference (see “Methods”). The p -values were generated by two-sided Student’s t -test with Benjamini-Hochberg multiple testing correction. Full results are provided in Supplementary Data 11. **b** MPRA enhancer activity at the 27 risk loci with at least one allelic enVar. Bar plots indicate the total number of variants at each locus. Variants with allelic enhancer activity (allelic enVars) are shown in red. Variants lacking allelic enhancer activity are shown in gray.

one ChIP-seq dataset (Supplementary Data 14), revealing groups of TFs and transcriptional regulators that allelically bind SLE-risk variants with genotype-dependent MPRA enhancer activity. For example, the rs3101018 variant, which is associated with SLE³⁸ and rheumatoid arthritis³⁹ in Europeans, shows 1.7-fold stronger enhancer activity for the reference/non-risk ‘C’ allele compared to the non-reference/risk ‘T’ allele (Fig. 4a). These results are consistent with a previously established eQTL obtained from GTEx⁴⁰, which demonstrates higher Complement C4A (*C4A*) expression in EBV-transformed B cell lines for the rs3101018 ‘C’ allele than the ‘T’ allele (Fig. 4b). Our MARIO allelic ChIP-seq analysis reveals 15 regulatory proteins that prefer the ‘C’ allele and 2 that prefer the ‘T’ allele (Fig. 4c and Supplementary Fig. 4a). Among these, particularly robust signal is obtained for ATF7, with one experimental replicate in GM12878 displaying 77 vs. 18 reads (‘C’ vs. ‘T’) and another showing 66 vs. 23 reads (‘C’ vs. ‘T’) (Supplementary Data 14). Moreover, CREB1 and CREM strongly favor the ‘C’ allele as well (Supplementary Data 14). In agreement with these data, computational analysis of the DNA sequences surrounding this variant predicts that ATF7, CREB1, and CREM will all bind more strongly to the ‘C’ than the ‘T’ allele (Fig. 4d). Intriguingly, ten additional proteins (FOXK2, PKNOX1, ARID3A, ZBTB40, ZNF217, ARNT, ELF1, IKZF2, MEF2B, and FOXM1) also bind allelically and have known DNA-binding motifs⁴¹, but none of them have binding sites altered by the variant. Further, we do not observe allelic chromatin accessibility in an available GM12878 ATAC-seq dataset (9 vs. 7 unique reads). Together, these results reveal a potentially causative SLE regulatory mechanism involving weaker direct binding of ATF7/CREB1/CREM to the ‘T’ risk allele, altering the recruitment of additional proteins to the locus and lowering the expression of *C4A*.

We observe a similar phenomenon for the rs2069235 variant, which is associated with SLE in Asian ancestries⁴² and rheumatoid arthritis in Europeans⁴³. rs2069235 displays much stronger enhancer activity for the ‘A’ (non-reference/risk) allele compared to the ‘G’ (reference/non-risk) allele (Fig. 4e), consistent with the established synaptogyrin 1 (*SYNGR1*) eQTL in

EBV-transformed B cell lines⁴⁰ (Fig. 4f). Inspection of our allelic ChIP-seq data reveals 14 proteins preferentially binding the ‘A’ allele, with none preferring the ‘G’ allele (Fig. 4g and Supplementary Fig. 4b). Among these 14 proteins, only ELF1 has its binding site directly altered by the variant (Fig. 4h). Strikingly, 55 of the 78 available H3K27ac datasets are allelic at this variant, with all 55 preferring the ‘A’ allele. Likewise, 24 of 46 H3K4me1 datasets are allelic, with all of them also preferring the ‘A’ allele. Both of these histone marks are indicative of active chromatin²⁸. Only a single histone mark dataset prefers the ‘G’ allele—the H3K27me3 mark, which is indicative of silenced chromatin²⁸ (Fig. 4g and Supplementary Fig. 4b). Together, these data are consistent with a potentially causative SLE molecular mechanism involving an allele-dependent enhancer consisting of stronger direct binding of ELF1 to the ‘A’ risk allele, along with indirectly altered binding of multiple additional TFs to this locus.

Genotype-dependent binding to SLE variants with allelic enhancer activity by variant overlapping and variant adjacent TFs.

As illustrated by the above examples, a particular TF can be involved in allelic mechanisms that are either directly impacted by a given variant (i.e., the variant directly alters the TF’s DNA-binding site) or indirectly impacted by the variant (i.e., the variant alters the DNA binding of the TF’s physical interaction partner, modulates chromatin accessibility, or affects another mechanism). At a given locus, we designate such TFs as variant overlapping and variant adjacent TFs, respectively (Fig. 5a). We next sought to discover such TFs at the 51 allelic enVars. At each allelic enVar locus, we identified variant overlapping TFs as those TFs predicted to have strong binding to one allele and weak binding to the other allele. Likewise, we identified variant adjacent TFs as those TFs with proximal strong predicted binding sites that do not directly overlap the variant (see “Methods”). We then searched for particular TFs that tend to act as variant overlapping TFs or as variant adjacent TFs at the 51 allelic enVars using a proportion test (see “Methods”) and confirmed that their binding site locations are distributed relative to the variant as expected

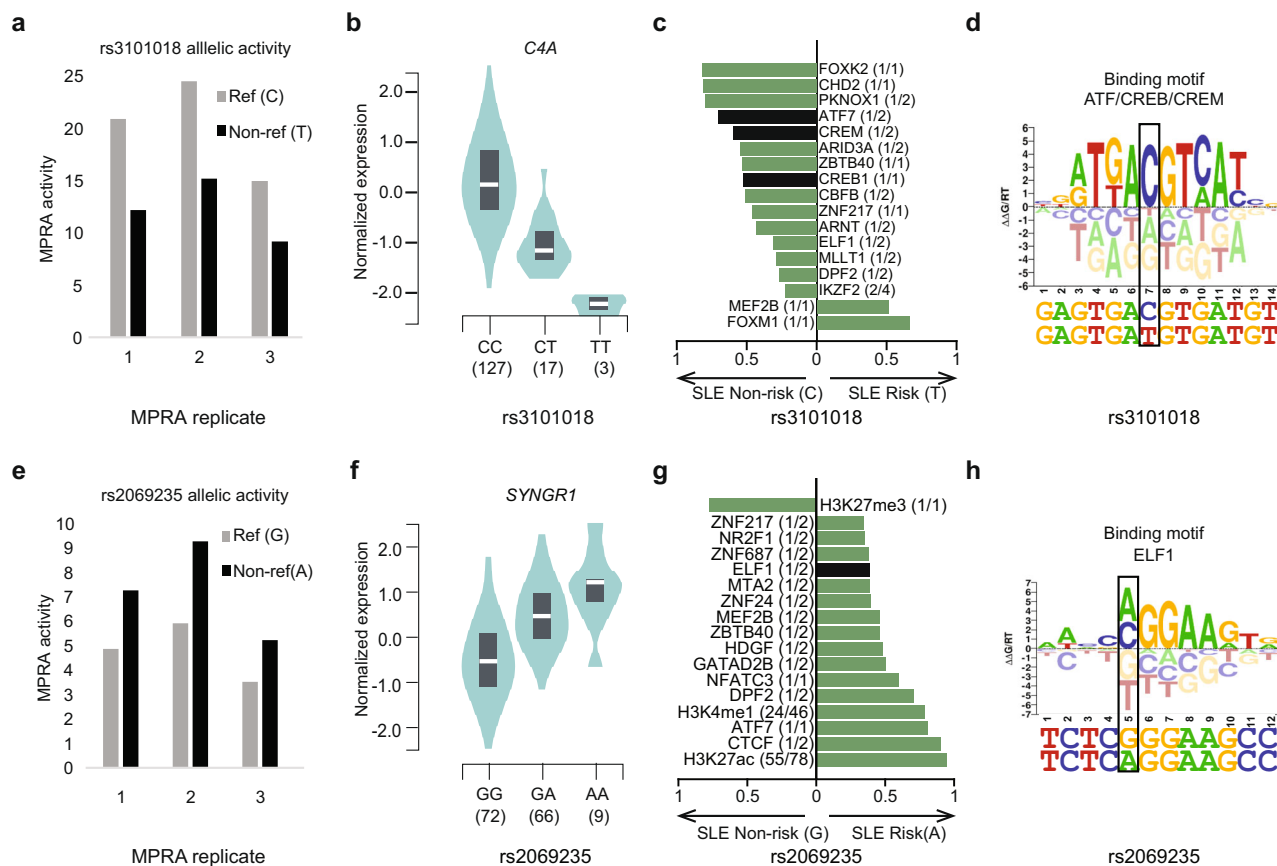


Fig. 4 Lupus risk allele-dependent gene regulatory mechanisms at the *C4A* and *SYNGR1* genomic loci. **a, e** Normalized MPRA enhancer activity of each experimental replicate for rs3101018 and rs26069235. **b, f** Expression trait quantitative loci (eQTLs) revealing genotype-dependent expression of *C4A* and *SYNGR1* for rs3101018 (CC, $n = 127$ biologically independent samples; CT, $n = 17$; TT, $n = 3$) and rs26069235 (GG, $n = 72$ biologically independent samples; GA, $n = 66$; AA, $n = 9$) in EBV-transformed B cell lines (GTEx). **c, g** Genotype-dependent activity of transcription factors, transcriptional regulators, and histone marks in EBV-transformed B cell lines for rs3101018 and rs26069235. Results with MARIO ARS value >0.4 and consistent allelic imbalance across ChIP-seq datasets are included (see “Methods”). The X-axis indicates the preferred allele, along with a value indicating the strength of the allelic behavior, calculated as one minus the ratio of the weak to strong read counts (e.g., 0.5 indicates the strong allele has twice the reads of the weak allele). The median value is plotted when data from multiple cell lines are available, with full results provided in Supplementary Fig. 4. The numbers in parentheses represent the number of ChIP-seq datasets with significant allelic activity (i.e., MARIO ARS value >0.4) out of the number of datasets where the given variant is inside a ChIP-seq peak and is also heterozygous in the given cell line. Variant overlapping TFs are indicated in black. Variant adjacent TFs are shown in green (see definition in Fig. 5a). **d, h** DNA-binding motif logos are shown for the ATF/CREB/CREM family, and ELF1 in the context of the DNA sequence surrounding rs3101018 and rs26069235, respectively. Tall nucleotides above the X-axis indicate preferred DNA bases. Bases below the X-axis are disfavored. In **(b)** and **(f)**, data are represented as a violin plot where the middle line is the median, the lower and upper hinges correspond to the first and third quartiles, with the rotated kernel density plot shown on each side. The data used for the analyses were obtained from the Genotype-Tissue Expression (GTEx) Portal on 11/12/2020. The GTEx Project was supported by the Common Fund of the Office of the Director of the National Institutes of Health, and by NCI, NHGRI, NHLBI, NIDA, NIMH, and NINDS.

(Fig. 5b). Consistent with our results at the *C4A* and *SYNGR1* loci, variant overlapping TFs include members of the ETS (e.g., ELF1) and ATF-like (e.g., ATF7) families, along with other TFs whose genetic loci are associated with SLE, including IRF5⁴⁴ (Fig. 5c and Supplementary Data 15). Variant adjacent TFs represent a distinct class, but also include several TFs with SLE genetic associations, including NFkB⁴⁵, the Ikaros (IKZF) family⁴⁶, and HMGA family members⁴⁷ (Fig. 5d and Supplementary Data 15). Collectively, these analyses reveal two distinct classes of TFs at a given SLE-associated locus that both likely play key roles in SLE mechanisms, along with particular TFs that tend to participate in one class or the other.

Allelic transcription regulatory mechanisms shared and unique across cell types. To explore the cell-type specificity of allelic enhancer activity, we transfected our SLE MPRA library into Jurkat cells, a T cell line, as T cells are another important cell type

in SLE². We identified 92 SLE-risk variants as allelic enVars in Jurkat cells, 25% of which were also found in GM12878 (Supplementary Fig. 5a, b and Supplementary Data 11). We then repeated the experiment in Jurkat cells stimulated with the inflammatory cytokine TNF α , a key cytokine in SLE development⁴⁸, to identify stimulation-dependent allelic enVars. This resulted in the identification of 102 allelic enVars, 28 of which were specific to the stimulated Jurkat cells (Supplementary Fig. 5c, d and Supplementary Data 11). Altogether, our study identified a total of 145 allelic enVars across 50 independent SLE-risk loci (Supplementary Fig. 5e). These results highlight allelic transcriptional regulatory mechanisms that are both cell-type and inflammatory signaling-dependent.

In summary, through the application of an allelic MPRA library to the EBV-transformed B cell line GM12878, we identified global transcriptional enhancer activity at 16% of SLE-associated genetic variants (enVars), with particular

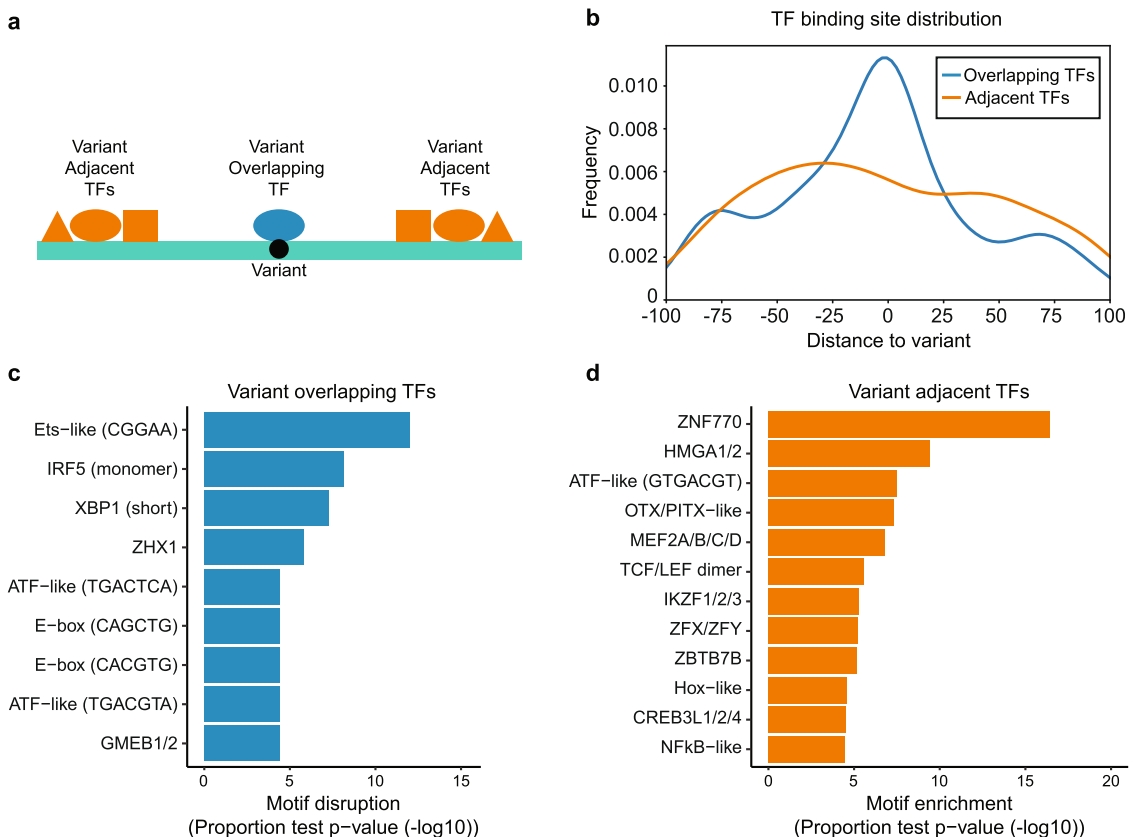


Fig. 5 Identification of variant overlapping and variant adjacent TFs. **a** Model of variant overlapping and variant adjacent transcription factors (TFs). Variant overlapping TFs (blue) allelically bind on top of variants, while variant adjacent TFs (orange) allelically bind near variants. **b** TF binding site location distribution for variant overlapping (blue) and variant adjacent (orange) TFs, relative to allelic enVars. **c** TF motif families enriched for participating as variant overlapping TFs at allelic enVars. Motif disruption *p*-values were estimated by a two-sided proportions test by comparing the fraction of motif disruption events at allelic enVars to the fraction observed at non-allelic enVars (see “Methods”). **d** TF motif families enriched for participating as variant adjacent TFs at allelic enVars. Motif enrichment *p*-values were estimated by a two-sided proportions test by comparing the fraction of predicted TF binding sites in allelic enVars to random expectation (see “Methods”). For both the variant overlapping and variant adjacent analyses, motif families are shown with $p_{adj} < 0.0001$ and three or more allelic events at allelic enVar loci, or five or more predicted binding sites at allelic enVar loci, respectively.

transcriptional regulatory proteins concentrated at these genomic locations. We further identified 51 SLE-risk variants with allelic enhancer activity (allelic enVars) that we now nominate as plausibly causal by acting through genotype-dependent changes to enhancer activity in GM12878. Upon comparison to allelic enhancer activity in the Jurkat T cell line, we identified shared and unique allelic transcriptional regulatory mechanisms at SLE-risk loci. Using experimental TF ChIP-seq data and TF binding site motif scanning, we propose a model where the collective action of the genotype-dependent binding of particular variant overlapping and variant adjacent TFs leads to genotype-dependent transcriptional activity at SLE-risk loci.

Discussion

Genome-wide association studies identify genetic loci with statistical disease associations. However, each risk locus often contains many plausibly causal variants due to linkage disequilibrium. This study is the first direct genome-wide measurement of enhancer activity at the ~3000 known SLE genetic risk variants in any context. Unbiased experimental approaches such as MPRA are vital for resolving causal variants and their molecular mechanisms of action.

Our results indicate that 16% of the SLE-risk variants examined in this study have enhancer activity in the EBV-transformed B cell line GM12878. Furthermore, 51 of these enhancer variants

at 27 loci have allelic enhancer activity. These findings are consistent with the theory that a large proportion of the genetic risk of SLE is mediated through transcriptional perturbation of critical B cell genes. Importantly, SLE-risk loci exhibit cell-type and inflammatory signaling-dependent allelic enhancer activity, with only 25% of allelic enVars shared between GM12878 and Jurkat cells. These results highlight both shared and unique allelic transcriptional regulatory mechanisms for SLE risk, which underlines the importance of the cell-type and cell-state in which the MPRA is performed.

In this study, we used the EBV-transformed B cell line GM12878 as a model for exploring the effects of SLE-risk variants. Previous studies have shown that B cells play a critical role in SLE development as immune cells that secrete autoantibodies driving etiology¹¹. The relationship between EBV and SLE is widely appreciated. For example, EBV-infected B cells are more prevalent in SLE patients than in healthy people^{13,14} and patients with SLE have a higher EBV viral load and infection rate relative to controls^{15,16}. In addition, the EBV transcriptional regulator EBNA2 occupies SLE-risk loci in a genotype-dependent manner¹⁰. In vivo, EBV infection can convert primary B cells to activated lymphoblasts^{49,50}. EBV will eventually enter latency in resting memory B cells and establish lifelong infection⁵¹. In vitro, EBV infection transforms B cells into immortalized lymphoblastoid cell lines (LCLs)¹⁷. While we chose the EBV-transformed B cell line GM12878 as the primary disease model for our study,

there are limitations to the use of EBV-transformed B cell lines. For example, GM12878 is an immortalized cell line with differences in DNA methylation and gene expression levels from resting and activated B cells^{52,53}. A recent study also suggests that EBV infection causes B cells to undergo a germinal center-like differentiation into cells partially resembling plasmablasts and early plasma cells⁵⁴, which is only a transient stage *in vivo*. Altogether, these limitations need to be considered when interpreting allelic transcriptional regulation of SLE-risk variants in EBV-transformed B cell lines.

A critical finding of this study is that SLE-risk variants with allelic enhancer activity likely alter the binding of many TFs. Although variants can directly affect the binding of variant overlapping TFs via disruption of a DNA-binding site, they can also simultaneously alter the binding of other variant adjacent TFs, presumably via genomic mechanisms such as altered chromatin accessibility, altered histone marks, indirect TF recruitment through physical interactions, changes in DNA shape, or changes to protein interaction partner DNA binding. This finding corroborates the previously proposed genetic variation-mediation model of motif-dependent and motif-independent TF binding^{55–57}. In general, a given TF can be variant overlapping at one locus and variant adjacent at another, as exemplified by ATF7 (Fig. 4c, g). Nonetheless, particular TF families tend to act as variant overlapping TFs at SLE loci (such as Ets, E-box, and ATF), whereas others tend to act as variant adjacent TFs (such as HMGA, Hox, and NFκB). Notably, many of these variant overlapping and variant adjacent TFs are themselves encoded by genetic risk loci associated with SLE (e.g., IRF5⁴⁴, NFκB⁴⁵, and ETS1⁵⁸), suggesting that there are multiple means through which a particular TF can contribute to disease-based genetic mechanisms. For example, IRF5 targets might be mis-regulated in an SLE patient due to genetic associations in the promoter of *IRF5* that result in altered IRF5 protein levels⁴⁴, or by genetic variants located within or adjacent to IRF5 binding sites at other genomic loci. It is currently unknown if these TF attributes are shared with other human diseases.

This study reveals possible causal genetic mechanisms involving altered binding of particular TFs at two important SLE-risk loci. *C4A* is a component of the inflammatory complement pathway that is critical for the appropriate clearance of apoptotic cells⁵⁹. People without *C4A* due to rare, protein-changing mutations are at a greatly increased risk for autoimmune diseases, including type I diabetes and SLE⁶⁰. Further, the risk of developing SLE is 2.62 times higher in subjects with low total *C4A*⁵⁹. Consistent with this observation, the SLE-risk allele of rs3101018 at this genetic locus identified in our MPRA is associated with lower *C4A* expression (Fig. 4a). Moreover, this variant is an eQTL for *C4A* in EBV-transformed B cell lines and whole blood cells, with the risk allele displaying lower *C4A* expression⁶¹ (Fig. 4b and Supplementary Data 7). rs3101018 is located in the Human Leukocyte Antigen (HLA) region of the genome. While many genetic variants at this locus alter amino acid usage in major histocompatibility complex molecules and affect antigen presentation, non-coding genetic variants across the HLA region have also been demonstrated to affect gene expression independently of HLA-type^{62–64}. In addition, the SLE-risk locus encoding *SYNGRI* was recently identified in a high-density genotyping study of subjects with Asian ancestry⁴² and also increases disease risk for schizophrenia⁶⁵, primary biliary cirrhosis⁶⁶, and rheumatoid arthritis⁶⁷. *SYNGRI* is an integral membrane protein that is most robustly expressed in neurons of the central nervous system; however, there is measurable transcription and translation of *SYNGRI* in other tissues, including developing B cells⁶⁸. eQTL data from EBV-transformed B cell lines and whole blood cells⁶¹ (Fig. 4f and Supplementary Data 7) further support our

MPRA-based findings of SLE-risk genotype-dependent enhancer activity and gene expression at this locus. Altogether, the results of our MPRA study provide mechanistic insight into these variants through the identification of allelic eVars that facilitate SLE-risk genotype-dependent gene expression.

GWAS provides important discernment of the genetic origins of disease. In conjunction with other genome-scale assays such as ATAC-seq, ChIP-seq, and HiChIP-seq, MPRA reveals likely causal variants and genes, and enables the assembly of causal mechanisms affecting gene expression. In this study, we used MPRA to uncover specific genetic variants within the risk haplotypes of a complex disease in specific cell types. Our integrative analyses reveal specific molecular mechanisms underlying genotype-dependent transcriptional regulation and SLE disease risk. We conclude that the MPRA is a robust tool for the nomination of causal genetic risk variants for any phenotype or disease with risk loci that act through genotype-dependent gene regulatory mechanisms, with this study providing a blueprint for dissecting the genetic etiology of many complex human diseases.

Methods

Variation selection and DNA sequence generation. All SLE-associated genetic risk loci reaching genome-wide significance published through March 2018 were included in this study^{31,38,44,58,69–78}. A total of 91 genetic risk loci were used for linkage disequilibrium (LD) expansion ($r^2 > 0.8$) based on 1000 Genomes Data⁷⁹ in the ancestry(ies) of the initial genetic association using PLINK(v1.90b)⁸⁰ (Supplementary Data 1). All expanded variants were updated to the dbSNP 151 table⁸¹ from the UCSC table browser⁸² based on either variant name or genomic location. Unmappable variants were discarded. We also included 20 genetic variants from the Tewhey et al.¹⁹ study as positive and negative controls.

For single nucleotide polymorphisms, we pulled 170 base pairs (bps) of hg19-flanking DNA sequences for every allele, with the variant located in the center (84 bps upstream and 85 bps downstream of the variant). For the other types of variants (indels), we designed the flanking sequences to ensure that the longest allele has 170 bps. Adapters (15 bps) were added to each sequence at either end (5'-ACTGGCCGCTTGACG - [170 bp oligo] - CACTGCGGCTCCTGC-3') to make a 200 bp DNA sequence (Supplementary Data 4). For all resulting sequences, we created a forward and reverse complement sequence to compensate for possible DNA synthesis errors. A total of 12,478 oligos (3093 variants, 6239 alleles) were obtained from Twist Bioscience.

Library assembly. For assembly of the MPRA library, we followed the procedure described by Tewhey et al.¹⁹ with minor modifications. In brief, we first created the empty vector pGL4.23ΔxbaΔluc from pGL4.23[luc2/minP] using primer Q5_deletion_rev and Q5_deletion_fwd following the manufacturer's instruction of the Q5 Site-Directed Mutagenesis Kit. Then, 20 bps barcodes were added to the synthesized oligos through 24X PCR with 50 μL system, each containing 1.86 ng oligo, 25 μL NEBNext® Ultra™ II Q5® Master Mix, 1 μM MPRA_v3_F, and MPRA_v3_201_R. PCR was performed under the following conditions: 98 °C for 2 min, 12 cycles of (98 °C for 10 s, 60 °C for 15 s, 72 °C for 45 s), 72 °C for 5 min. Amplified product was purified and cloned into SfiI digested pGL4.23ΔxbaΔluc by Gibson assembly at 50 °C for 1 h. The assembled backbone library was purified and then transformed into *Escherichia coli* (*E. coli*) through electroporation (2 kV, 200 ohm, 25 μF). Electroporated *E. coli* was expanded in 200 mL of LB Broth buffer supplemented with 100 μg/mL of carbenicillin at 37 °C for 12 to 16 h. Plasmid was then extracted using the QIAGEN Plasmid Maxi Kit.

We next created the pGL4.23[eGFP/miniP] plasmid. An eGFP fragment was amplified from MS2-P65-HSF1_GFP (Addgene #61423) through PCR with a 50 μL system containing 1 ng plasmid, 25 μL NEBNext® Ultra™ II Q5® Master Mix, 0.5 μM GFP_seq_MS2-P65-HSF1_GFP_FWD, and GFP_seq_MS2-P65-HSF1_GFP_REV. PCR was performed under the following conditions: 98 °C for 2 min, 20 cycles of (98 °C for 10 s, 60 °C for 15 s, 72 °C for 30 s), 72 °C for 5 min. The amplified fragment was purified and then inserted into XbaI and NcoI digested pGL4.23[luc2/minP] through Gibson assembly at 50 °C for 1 h. The assembled plasmid was purified and then transformed into *E. coli* through chemical transformation. Transformed *E. coli* was expanded in 100 mL of LB Broth buffer supplemented with 100 μg/mL of carbenicillin at 37 °C for 12–16 h. Plasmid was then extracted using the QIAGEN Plasmid Maxi Kit.

A miniP + eGFP fragment was amplified from pGL4.23[eGFP/miniP] through 8X PCR with 50 μL system, each containing 1 ng plasmid, 25 μL NEBNext® Ultra™ II Q5® Master Mix, 0.5 μM 200-MPRA_v3_GFP_Fusion_v2_F, and 201-MPRA_v3_GFP_Fusion_v2_R. PCR was performed under the following conditions: 98 °C for 2 min, 20 cycles of (98 °C for 10 s, 60 °C for 15 s, 72 °C for 45 s), 72 °C for 5 min. The amplified product was purified and then inserted into AsiSI digested backbone library through Gibson assembly at 50 °C for 1.5 h to create the

transfection library. The resulting library was re-digested by RecBCD and AsiSI, purified, and then transformed into *E. coli* through electroporation (2 kV, 200 ohm, 25 μ F). Transformed *E. coli* was cultured in 5 L of LB Broth buffer supplemented with 100 μ g/mL of carbenicillin at 37 °C for 12–16 h. The plasmid was then extracted using the QIAGEN Endo-free Plasmid Giga Kit.

Sequencing library for oligo and barcode association. The oligo and barcode regions were amplified from the backbone library through 4X PCR with a 100 μ L system containing 200 ng plasmid, 50 μ L NEBNext® Ultra™ II Q5® Master Mix, 0.5 μ M TruSeq_Universal_Adapter_P5, and MPRA_v3_TruSeq_Amp2Sa_F_P7. PCR was performed under the following conditions: 95 °C for 20 s, 6 cycles of (95 °C for 20 s, 62 °C for 15 s, 72 °C for 30 s), 72 °C for 2 min. The product was then purified, and indices were added through a 100 μ L system containing all purified product, 50 μ L NEBNext® Ultra™ II Q5® Master Mix, 0.5 μ M TruSeq_Universal_Adapter_P5, and index primer. PCR was performed as above, except for only five cycles. Samples were purified, molar pooled, and sequenced using 2 \times 125 bp on Illumina NextSeq 500.

Transfection. The GM12878 cell line was grown in RPMI medium supplemented with 10% FBS, 100 units/mL of penicillin, and 100 μ g/mL of streptomycin. Cells were seeded at a density of 5×10^5 cells/mL the day before transfection. For triplicate transfections, we collected a total of 5×10^7 cells per replicate. Cells were then suspended with 50 μ g transfection library plasmid in 400 μ L Buffer R. Electroporation was performed with the Neon transfection system in 100 μ L tips with 3 pulses of 1200 V, 20 ms each. After transfection, cells were recovered in 50 mL pre-warmed RPMI medium supplemented only with 10% FBS for 24 h. Cells were then collected for preparation of the sequencing library for barcode counting.

The Jurkat cell line was grown in RPMI medium supplemented with 10% FBS, 100 units/mL of penicillin, and 100 μ g/mL of streptomycin. Cells were seeded at a density of 5×10^5 cells/mL the day before transfection. For each experimental group, we collected a total of 5×10^7 cells per replicate for 5 replicates. Cells were then resuspended with 50 μ g transfection library plasmid in 400 μ L Buffer R. Electroporation was performed with the Neon transfection system in 100 μ L tips with 3 pulses of 1350 V, 10 ms each. After transfection, cells were recovered in 50 mL pre-warmed RPMI medium supplemented only with 10% FBS for 24 h. After recovery, cells were supplemented with or without 100 ng/mL TNF α for 24 h. Cells were then collected for preparation of the sequencing library for barcode counting.

Sequencing library for barcode counting. For samples from GM12878 cells, total RNA of transfected cells was extracted by the RNeasy Midi Kit following the manufacturer's instruction. Extracted RNA was subjected to DNase treatment in a 375 μ L system with 2.5 μ L Turbo DNase and 37.5 μ L Turbo DNase Buffer at 37 °C for 1 h. 3.75 μ L 10% SDS and 37.5 μ L 0.5 M EDTA were added to stop DNase with 5 min of incubation at 75 °C. The whole volume was used for eGFP probe hybridization in an 1800 μ L system, with 450 μ L 20X SSC buffer, 900 μ L Formamide and 1 μ L of each 100 μ M Biotin-labeled GFP probe One to Three. The probe hybridization was performed through incubation at 65 °C for 2.5 h. 200 μ L Dynabeads™ MyOne™ Streptavidin C1 was prepared according to the manufacturer's instruction. The beads were suspended in 250 μ L 20X SSC Buffer and incubated with the above probe hybridization reaction at room temperature for 15 min. Beads were then collected on a magnet and washed with 1X SSC Buffer once, and 0.1X SSC Buffer twice. eGFP mRNA was eluted first through adding 12.5 μ L ddH₂O, heating at 70 °C for 2 min and collecting on a magnet, then adding another 12.5 μ L ddH₂O, heating at 80 °C for 2 min and collecting on a magnet. All collected elution was performed with another DNase treatment in a 30 μ L system containing 0.5 μ L Turbo DNase and 3 μ L Turbo DNase Buffer at 37 °C for 1 h. 0.5 μ L 10% SDS was added to halt DNase reaction. Eluted mRNA was purified through RNA Clean SPRI Beads. mRNA was reverse transcribed to cDNA using SuperScript™ IV First-Strand Synthesis System with gene specific primer MPRA_v3_Amp2Sc_R, following the manufacturer's instruction. cDNA and plasmid control were then used for building sequencing libraries following the Tag-seq Library Construction section in the paper of Tewhey et al.¹⁹. In brief, 1 μ L of cDNA and plasmid control samples were used to estimate the relative concentration of eGFP in the 10 μ L system containing 5 μ L NEBNext® Ultra™ II Q5® Master Mix, 0.6 μ L SYBR green I diluted 1:1000 (Life Technologies, S7563), and 0.5 μ M TruSeq_Universal_Adapter_P5 and MPRA_V3_Illumina_GFP_F. PCR was performed under the following conditions: 95 °C for 20 s, 40 cycles of (95 °C for 20 s, 65 °C for 20 s, 72 °C for 30 s), 72 °C for 2 min. According to the cycle threshold, all cDNA and plasmid control samples were diluted to match the sample with the lowest concentration. A total of two PCRs were needed for building the sequencing library. The first PCR was performed with 10 μ L of normalized samples in the 50 μ L system containing 25 μ L NEBNext® Ultra™ II Q5® Master Mix, 0.5 μ M TruSeq_Universal_Adapter_P5, and MPRA_V3_Illumina_GFP_F. PCR was performed under the following conditions: 95 °C for 20 s, corresponding cycles of (95 °C for 20 s, 65 °C for 20 s, 72 °C for 30 s), 72 °C for 2 min. The product was then purified, and indices were added through a 100 μ L system containing all purified product, 50 μ L NEBNext® Ultra™ II Q5® Master Mix, 0.5 μ M TruSeq_Universal_Adapter_P5, and index primer. PCR was performed as above, except for only 6 cycles. Samples were purified, molar pooled, and sequenced using 1 \times 75 bp on Illumina NextSeq 500.

For samples from Jurkat cells, total DNA and RNA of transfected cells were extracted by the Qiagen ALLPrep DNA/RNA Mini Kit following the manufacturer's instruction²³. Extracted RNA was processed the same as above to obtain cDNA. cDNA, extracted DNA, and plasmid control were then used for building sequencing libraries with the same protocol described above. Samples were purified, molar pooled, and sequenced using 1 \times 100 bp on Illumina NovaSeq 6000.

All primers used in this study are provided in Supplementary Table 1.

Oligo and barcode association. Paired-end, 125 bp reads were first quality filtered using Trimmomatic (v0.38)⁸³ (flags: PE -phred33, LEADING:25, TRAILING:25, MINLEN:80). Read 1 was then separated into the 20 bp barcode region and the oligo-matching region. The trimmed oligo-matching regions of Read 1 and Read 2 were mapped back to the synthesized oligo sequences using Bowtie2 (v2.3.4.1)⁸⁴ (flags: -X 250, -very-sensitive, -p 16). Barcodes were then associated with the oligo sequences using the read ID. Only uniquely mapped barcodes were used for downstream analysis.

Barcode counting. Single-end 75/100 bp reads were first quality filtered using Trimmomatic (v0.38)⁸³ (flags: PE -phred33, LEADING:3, TRAILING:3, MINLEN:70). Each read was then separated into the 20 bp barcode region and the constant region. The trimmed constant regions of the reads were mapped back to the constant region within the eGFP 3' UTR using Bowtie2 (v2.3.4.1)⁸⁴ (flags: -very-sensitive, -p 16). Only reads with Levenshtein distance of 4 or less within the constant region and perfect matches to the two bases directly adjacent to the barcode were kept. Barcodes were then associated with the retained reads using the read ID. Only barcodes that met our quality threshold requirements described above in the Methods section "Oligo and barcode association" were used for downstream analysis.

Enhancer variant (EnVar) identification. We followed the procedures described in the "Identification of Regulatory Oligos" section of Tewhey et al.¹⁹ with minor modifications. In brief, oligos (alleles) with 30 or more unique barcodes from the plasmid control were included for analysis. All barcodes were summarized at the oligo level. Barcode count totals for each oligo, including all SLE variants and the 20 control variants, were passed into DESeq2 (v1.28.1)⁸⁵ in R (v3.5.3) to estimate the fold change and significance between plasmid controls (Supplementary Note 1 and Supplementary Fig. 5f, g) and the experimental replicates. A Benjamini–Hochberg FDR adjusted *p*-value of <0.05 was required for significance. Only significant alleles with greater than or equal to a 1.5x fold change were identified as enhancer alleles (enAlleles). A variant was identified as an enhancer variant (enVar) if any allele of this variant was an enAllele. Results for the 20 control variants were compared to data from Tewhey et al.¹⁹ to estimate accuracy, sensitivity, and specificity.

Allelic enVar identification. Only enVars were considered for allelic analysis. The barcode counts from every allele of each enVar were used for calculating *p*-values by comparing the log₂ ratios of the non-reference allele vs the reference allele, normalized by plasmid controls, using Student's *t*-test^{19,35}. *p*-values were adjusted with the Benjamini–Hochberg FDR-based procedure. A corrected *p*-value of <0.05 was required for significance. Only significant alleles with 25%-fold changes or greater were identified as allelic enVars (Supplementary Note 2 and Supplementary Fig. 6). We have created an R package (mraprofiler) for performing this analysis, which is available on the Weirauch lab GitHub page (<https://github.com/WeirauchLab/mraprofiler>).

Gene annotation. We annotated each SLE genetic variant with its nearest gene using the NCBI RefSeq table⁸⁶ downloaded from the UCSC table browser⁸². enVars were annotated using a combination of DNA looping interactions (GM12878 Capture Hi-C data^{87,88}) and eQTL data obtained from the eQTL Catalog, a resource that contains quality-controlled, uniformly re-computed eQTLs from 19 eQTL publications^{61,89–108}, EBV-transformed B cell lines (GTEx Analysis V7 (dbGaP Accession phs000424.v7.p2))⁴⁰ and other individual studies^{109–112}. For all variants, the target genes were annotated (Supplementary Data 2) using the union of promoter interacting genes and eQTL genes from B cells with and without EBV transformation, when available. Otherwise, target genes were annotated as the nearest gene. Allelic enVar gene targets were classified into four tiers: a Tier (1) variant is both an eQTL and also loops to the promoter of the same gene; a Tier (2) variant has an eQTL for at least one gene; a Tier (3) variant only loops to the promoter of at least one gene; a Tier (4) variant is neither an eQTL nor loops to the promoter of any gene (Supplementary Data 12).

TF binding site motif enrichment analysis. To identify specific TFs whose binding might contribute to the enhancer activity observed in our MPRA experiments, we performed HOMER (v4.9)³³ TF binding site motif enrichment analysis. Specifically, we used HOMER to calculate the enrichment of each motif in the sequence of enAlleles compared to the sequences of non-enAlleles. HOMER was modified to use the large library of human position weight matrix (PWM) binding

site models contained in build 2.0 of the Cis-BP database³⁴ and a log base 2 likelihood scoring system.

GO enrichment analysis. Enrichr^{113,114} was used for GO enrichment analysis. In short, the target genes of enVars were passed to Enrichr for analysis. Results from the GO biological process (2018) category were used (Supplementary Data 8, Supplementary Fig. 3).

Identification and processing of publicly available LCL ChIP-seq data. 1058 ChIP-seq datasets were obtained from the Gene Expression Omnibus (GEO)¹¹⁵ using custom scripts that searched for ChIP-seq experiments performed in EBV-transformed lymphoblastoid cell lines (LCLs). The annotations for every dataset (assay type, cell line, assayed molecule) were manually checked by two authors (MTW and LCK) to ensure accuracy. The Sequence Read Archive (SRA) files obtained from GEO were analyzed using an automated pipeline. Briefly, the pipeline first runs QC on the FastQ files containing the sequencing reads using FastQC (v0.11.2). If FastQC detects adapter sequences, the pipeline runs the FastQ files through Trim Galore (v0.4.2)¹¹⁶, a wrapper script that runs cutadapt (v1.9.1)¹¹⁷ to remove the detected adapter sequence from the reads. The quality-controlled reads are then aligned to the reference human genome (hg19/GRCh37) using bowtie2 (v2.3.4.1)⁸⁴. The aligned reads (in .BAM format) are then sorted using samtools (v1.8.0)¹¹⁸ and duplicate reads are removed using picard (v1.89)¹¹⁹. Finally, peaks are called using MACS2 (v2.1.2) (flags: callpeak -g hs -q 0.01 -f BAM)⁴⁴. ENCODE blacklist regions¹²⁰ were removed from the peak sets using the hg19-blacklist.bed.gz file available at https://github.com/Boyle-Lab/Blacklist/tree/master/lists/Blacklist_v1. ChIP-seq datasets GSM1666207, GSM2748907, and GSM1599157 were removed due to the low number of cells used in the experiments.

Functional genomics dataset enrichment analysis with RELI. We used the RELI (v0.9)¹⁰ algorithm to identify genomic features (TF binding events, histone marks, etc.) that coincide with enVars. As input, RELI takes the genomic coordinates of enVars. RELI then systematically intersects these coordinates with one of the GM12878 ChIP-seq datasets, and the number of input regions overlapping the peaks of this dataset (by at least one base) is counted. Next, a *p*-value describing the significance of this overlap is estimated using a simulation-based procedure. To this end, a ‘negative set’ is created for comparison to the input set, which in this study contains the set of non-enVars (i.e., variants with no allele having an adjusted *p*-value of <0.05 and more than 10%-fold change in the DESeq2 result). A distribution of expected overlap values is then created from 2000 iterations of randomly sampling from the negative set, each time choosing a set of negative examples that match the input set in terms of the total number of genomic loci. The distribution of the expected overlap values from the randomized data resembles a normal distribution and can thus be used to generate a Z-score and corresponding *p*-value estimating the significance of the observed number of input regions that overlap each ChIP-seq dataset.

We performed similar RELI analysis for allelic enVars. As input, we used the allelic enVar sites. For the ‘negative set’, we used the set of common SNPs taken from the dbSNP142 database downloaded from the UCSC table browser⁸².

Identification of allelic ChIP-seq reads using MARIO. To identify possible mechanisms underlying our allelic enVars, we applied our MARIO (v3.93)¹⁰ method to the LCL ChIP-seq dataset collection described above. In brief, MARIO identifies common genetic variants that are (1) heterozygous in the assayed cell line and (2) located within a peak in a given ChIP-seq dataset. It then examines the sequencing reads that map to each heterozygote in each peak for imbalance between the two alleles. Results are combined across experimental replicates to produce a robust Allelic Reproducibility Score (ARS). Results with MARIO ARS value >0.4 that also pass the following three post-processing filters were considered allelic. (1) The variant must be significantly allelic for a given protein/histone mark (ARS > 0.4) in at least 50% of the datasets in which that variant was heterozygous; (2) The same allele must be significantly preferred (ARS > 0.4) in at least 75% of the datasets where that variant shows significant allelic behavior; and (3) The replicates of a given experiment must all prefer the same strong allele. These post-processing filters were applied to remove results with inconsistent allelic imbalance, extending the procedures of our previous study¹⁰.

Identification of variant overlapping and variant adjacent TFs. Variant overlapping TFs were identified using an algorithm that compares predicted TF binding motif scores between the different alleles of each allelic enVar. First, we padded each allele of a given allelic enVar with 25 bps of upstream and downstream DNA sequence (a sufficient length to account for any known human TF binding sites¹²¹). The algorithm consists of two major components: (1) individually scoring the two alleles of a given variant with a given TF model; and (2) quantifying the difference in the binding intensity between these two alleles. DNA sequences are scored using the large collection of human TF position weight matrix (PWM) models contained in the Cis-BP database³⁴ and the log-likelihood PWM scoring system¹²². Since log-likelihood score distributions vary substantially (depending on the information

content of a given motif), we employ a simple scaled scoring system that maps a given log-likelihood score to the percentage of the maximum achievable log-likelihood score of the given motif—we refer to this value as the “relative PWM score”. We identify binding site altering events (i.e., “creating” or “breaking” a predicted binding site for a given TF motif) as cases where one allele has a relative PWM score of 70% or higher, and the other allele has a score of <40%. For a given variant, any TF with allelic ChIP-seq sequencing reads (see above) and a binding site altering event for any of its motifs was deemed a variant overlapping TF. Any TF with allelic ChIP-seq sequencing reads and a lack of a binding site altering event for any of its motifs was deemed a variant adjacent TF.

We next sought to identify particular TFs that tend to be variant overlapping TFs at SLE allelic enVars. To this end, we calculated the fraction of times each TF motif has a binding site altering event (as defined above) at SLE allelic enVars. As background, we calculated the fraction of times each TF motif has a binding site altering event at non-allelic enVars. The significance of the difference between these two fractions was then calculated using a proportions test. Results are provided in Supplementary Data 15.

We used a similar procedure to identify particular TFs that tend to be variant adjacent TFs at SLE allelic enVars. We performed HOMER (v4.9) motif enrichment analysis using the full 170 bp allelic enVar DNA sequences as input. The dinucleotide scrambled version of input sequences were used as background. The fractions of motif “hits” obtained in the foreground vs. background set were then compared, and significance was again calculated using a proportions test. Results are provided in Supplementary Data 15.

Reporting summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

All sequencing data that support the findings of this study are available in the Gene Expression Omnibus (GEO) database under accession number GSE143792. Full datasets and processed results are provided in the Supplementary Data. All other relevant data are available from the corresponding author upon request.

Code availability

Source code, with full documentation and examples, are freely available under the GNU General Public License on the WeirauchLab GitHub page: <https://github.com/WeirauchLab/mpraprofiler>. Additional modified scripts can be accessed upon request.

Received: 18 January 2020; Accepted: 16 February 2021;

Published online: 12 March 2021

References

- Carter, E. E., Barr, S. G. & Clarke, A. E. The global burden of SLE: prevalence, health disparities and socioeconomic impact. *Nat. Rev. Rheumatol.* **12**, 605–620 (2016).
- Tsokos, G. C. Systemic lupus erythematosus. *N. Engl. J. Med.* **365**, 2110–2121 (2011).
- Deng, Y. & Tsao, B. P. Genetic susceptibility to systemic lupus erythematosus in the genomic era. *Nat. Rev. Rheumatol.* **6**, 683–692 (2010).
- Visscher, P. M. et al. 10 Years of GWAS discovery: biology, function, and translation. *Am. J. Hum. Genet.* **101**, 5–22 (2017).
- Fike, A. J., Elcheva, I. & Rahman, Z. S. M. The post-GWAS era: how to validate the contribution of gene variants in lupus. *Curr. Rheumatol. Rep.* **21**, 3 (2019).
- Corradin, O. et al. Modeling disease risk through analysis of physical interactions between genetic variants within chromatin regulatory circuitry. *Nat. Genet.* **48**, 1313–1320 (2016).
- Farh, K. K. et al. Genetic and epigenetic fine mapping of causal autoimmune disease variants. *Nature* **518**, 337–343 (2015).
- Teruel, M. & Alarcon-Riquelme, M. E. The genetic basis of systemic lupus erythematosus: What are the risk factors and what have we learned. *J. Autoimmun.* **74**, 161–175 (2016).
- Hu, X. et al. Integrating autoimmune risk loci with gene-expression data identifies specific pathogenic immune cell subsets. *Am. J. Hum. Genet.* **89**, 496–506 (2011).
- Harley, J. B. et al. Transcription factors operate across disease loci, with EBNA2 implicated in autoimmunity. *Nat. Genet.* **50**, 699–707 (2018).
- Karrar, S., Cunninghame & Graham, D. S. Abnormal B cell development in systemic lupus erythematosus: what the genetics tell us. *Arthritis Rheumatol.* **70**, 496–507 (2018).
- Dorner, T., Giesecke, C. & Lipsky, P. E. Mechanisms of B cell autoimmunity in SLE. *Arthritis Res. Ther.* **13**, 243 (2011).

13. Draborg, A. H., Duus, K. & Houen, G. Epstein-Barr virus and systemic lupus erythematosus. *Clin. Dev. Immunol.* **2012**, 370516 (2012).
14. McClain, M. T. et al. Early events in lupus humoral autoimmunity suggest initiation through molecular mimicry. *Nat. Med.* **11**, 85–89 (2005).
15. James, J. A. et al. An increased prevalence of Epstein-Barr virus infection in young patients suggests a possible etiology for systemic lupus erythematosus. *J. Clin. Invest.* **100**, 3019–3026 (1997).
16. Hanlon, P., Avenell, A., Aucott, L. & Vickers, M. A. Systematic review and meta-analysis of the sero-epidemiological association between Epstein-Barr virus and systemic lupus erythematosus. *Arthritis Res. Ther.* **16**, R3 (2014).
17. Pope, J. H. Establishment of cell lines from peripheral leucocytes in infectious mononucleosis. *Nature* **216**, 810–811 (1967).
18. Patwardhan, R. P. et al. Massively parallel functional dissection of mammalian enhancers in vivo. *Nat. Biotechnol.* **30**, 265–270 (2012).
19. Tewhey, R. et al. Direct identification of hundreds of expression-modulating variants using a multiplexed reporter assay. *Cell* **165**, 1519–1529 (2016).
20. Ulirsch, J. C. et al. Systematic functional dissection of common genetic variation affecting red blood. *Cell Traits* **Cell** **165**, 1530–1545 (2016).
21. Liu, Y. et al. Functional assessment of human enhancer activities using whole-genome STARR-sequencing. *Genome Biol.* **18**, 219 (2017).
22. Vockley, C. M. et al. Massively parallel quantification of the regulatory effects of noncoding genetic variation in a human cohort. *Genome Res.* **25**, 1206–1214 (2015).
23. Klein, J. C. et al. Functional testing of thousands of osteoarthritis-associated variants for regulatory activity. *Nat. Commun.* **10**, 2434 (2019).
24. Choi, J. et al. Massively parallel reporter assays of melanoma risk variants identify MX2 as a gene promoting melanoma. *Nat. Commun.* **11**, 2718 (2020).
25. Lin, X. et al. FAM13A represses AMPK activity and regulates hepatic glucose and lipid metabolism. *iScience* **23**, 100928 (2020).
26. Madan, N. et al. Functionalization of CD36 cardiovascular disease and expression associated variants by interdisciplinary high throughput analysis. *PLoS Genet.* **15**, e1008287 (2019).
27. Zhang, S. et al. Allele-specific open chromatin in human iPSC neurons elucidates functional disease variants. *Science* **369**, 561–565 (2020).
28. Zhou, V. W., Goren, A. & Bernstein, B. E. Charting histone modifications and the functional organization of mammalian genomes. *Nat. Rev. Genet.* **12**, 7–18 (2011).
29. Ruer-Laventie, J. et al. Overexpression of Fkbp11, a feature of lupus B cells, leads to B cell tolerance breakdown and initiates plasma cell differentiation. *Immun. Inflamm. Dis.* **3**, 265–279 (2015).
30. Armstrong, D. L. et al. GWAS identifies novel SLE susceptibility genes and explains the association of the HLA region. *Genes Immun.* **15**, 347–354 (2014).
31. Han, J. W. et al. Genome-wide association study in a Chinese Han population identifies nine new susceptibility loci for systemic lupus erythematosus. *Nat. Genet.* **41**, 1234–1237 (2009).
32. Yang, J. et al. ELF1 is associated with systemic lupus erythematosus in Asian populations. *Hum. Mol. Genet.* **20**, 601–607 (2011).
33. Heinz, S. et al. Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Mol. Cell* **38**, 576–589 (2010).
34. Weirauch, M. T. et al. Determination and inference of eukaryotic transcription factor sequence specificity. *Cell* **158**, 1431–1443 (2014).
35. Ray, J. P. et al. Prioritizing disease and trait causal variants at the TNFAIP3 locus using functional and genomic features. *Nat. Commun.* **11**, 1237 (2020).
36. Stergachis, A. B. et al. Exonic transcription factor binding directs codon choice and affects protein evolution. *Science* **342**, 1367–1372 (2013).
37. Zhao, B. et al. Epstein-Barr virus exploits intrinsic B-lymphocyte transcription programs to achieve immortal cell growth. *Proc. Natl Acad. Sci. USA* **108**, 14902–14907 (2011).
38. International Consortium for Systemic Lupus Erythematosus G. Genome-wide association scan in women with systemic lupus erythematosus identifies susceptibility variants in ITGAM, PXX, KIAA1542 and other loci. *Nat. Genet.* **40**, 204–210 (2008).
39. Black, M. H. & Watanabe, R. M. A principal-components-based clustering method to identify multiple variants associated with rheumatoid arthritis and arthritis-related autoantibodies. *BMC Proc.* **3**, S129 (2009).
40. Consortium, G. T. The genotype-tissue expression (GTEx) project. *Nat. Genet.* **45**, 580–585 (2013).
41. Lambert, S. A. et al. Similarity regression predicts evolution of transcription factor sequence specificity. *Nat. Genet.* **51**, 981–989 (2019).
42. Sun, C. et al. High-density genotyping of immune-related loci identifies new SLE risk variants in individuals with Asian ancestry. *Nat. Genet.* **48**, 323–330 (2016).
43. Okada, Y. et al. Genetics of rheumatoid arthritis contributes to biology and drug discovery. *Nature* **506**, 376–381 (2014).
44. Kottyan, L. C. et al. The *IRF5-TNPO3* association with systemic lupus erythematosus has two components that other autoimmune disorders variably share. *Hum. Mol. Genet.* **24**, 582–596 (2015).
45. Li, Y. et al. Genetic variants of IkappaB kinase beta (IKKB) and polymerase beta (POLB) were not associated with systemic lupus erythematosus risk in a Chinese Han population. *PLoS ONE* **10**, e0132556 (2015).
46. Zhang, Y. M. et al. Association of the IKZF1 5' UTR variant rs1456896 with lupus nephritis in a northern Han Chinese population. *Scand. J. Rheumatol.* **46**, 210–214 (2017).
47. Patel, Z. H. et al. A plausibly causal functional lupus-associated risk variant in the STAT1-STAT4 locus. *Hum. Mol. Genet.* **27**, 2392–2404 (2018).
48. Zhu, L. J., Yang, X. & Yu, X. Q. Anti-TNF-alpha therapies in systemic lupus erythematosus. *J. Biomed. Biotechnol.* **2010**, 465898 (2010).
49. Zhou, H. et al. Epstein-Barr virus oncoprotein super-enhancers control B cell growth. *Cell Host Microbe* **17**, 205–216 (2015).
50. Kurth, J. et al. EBV-infected B cells in infectious mononucleosis: viral strategies for spreading in the B cell compartment and establishing latency. *Immunity* **13**, 485–495 (2000).
51. Babcock, G. J., Hochberg, D. & Thorley-Lawson, D. A. The expression pattern of Epstein-Barr virus latent genes in vivo is dependent upon the differentiation stage of the infected B cell. *Immunity* **13**, 497–506 (2000).
52. Hernando, H. et al. The B cell transcription program mediates hypomethylation and overexpression of key genes in Epstein-Barr virus-associated proliferative conversion. *Genome Biol.* **14**, R3 (2013).
53. Hansen, K. D. et al. Large-scale hypomethylated blocks associated with Epstein-Barr virus-induced B-cell immortalization. *Genome Res.* **24**, 177–184 (2014).
54. Mrozek-Gorska, P. et al. Epstein-Barr virus reprograms human B lymphocytes immediately in the prelatency phase of infection. *Proc. Natl Acad. Sci. USA* **116**, 16046 (2019).
55. Deplancke, B., Alpern, D. & Gardeux, V. The genetics of transcription factor DNA binding variation. *Cell* **166**, 538–554 (2016).
56. Kilpinen, H. et al. Coordinated effects of sequence variation on DNA binding, chromatin structure, and transcription. *Science* **342**, 744–747 (2013).
57. Reddy, T. E. et al. Effects of sequence variation on differential allelic transcription factor occupancy and gene expression. *Genome Res.* **22**, 860–869 (2012).
58. Lu, X. et al. Lupus risk variant increases pSTAT1 binding and decreases ETS1 expression. *Am. J. Hum. Genet.* **96**, 731–739 (2015).
59. Tsang, A. S. M. W. P. et al. Comprehensive approach to study complement C4 in systemic lupus erythematosus: gene polymorphisms, protein levels and functional activity. *Mol. Immunol.* **92**, 125–131 (2017).
60. Juptner, M. et al. Low copy numbers of complement C4 and homozygous deficiency of C4A may predispose to severe disease and earlier disease onset in patients with systemic lupus erythematosus. *Lupus* **27**, 600–609 (2018).
61. Momozawa, Y. et al. IBD risk loci are enriched in multigenic regulatory modules encompassing putative causative genes. *Nat. Commun.* **9**, 2427 (2018).
62. D'Antonio, M. et al. Systematic genetic analysis of the MHC region reveals mechanistic underpinnings of HLA type associations with disease. *Elife* **8**, e48476 (2019).
63. Lam, T. H., Shen, M., Tay, M. Z. & Ren, E. C. Unique allelic eQTL clusters in human MHC haplotypes. *G3* **7**, 2595–2604 (2017).
64. Aguiar, V. R. C., Cesar, J., Delaneau, O., Dermitzakis, E. T. & Meyer, D. Expression estimation and eQTL mapping for HLA genes with a personalized pipeline. *PLoS Genet.* **15**, e1008091 (2019).
65. Iatropoulos, P. et al. Association study and mutational screening of SYNGRI1 as a candidate susceptibility gene for schizophrenia. *Psychiatr. Genet.* **19**, 237–243 (2009).
66. Kim, K. et al. High-density genotyping of immune loci in Koreans and Europeans identifies eight new rheumatoid arthritis risk loci. *Ann. Rheum. Dis.* **74**, e13 (2015).
67. Liu, J. Z. et al. Dense fine-mapping study identifies new susceptibility loci for primary biliary cirrhosis. *Nat. Genet.* **44**, 1137–1141 (2012).
68. Wu, C. et al. BioGPS: an extensible and customizable portal for querying and organizing gene annotation resources. *Genome Biol.* **10**, R130 (2009).
69. Liu, L. et al. Genome-wide association study identifies three novel susceptibility loci for systemic lupus erythematosus in Han Chinese. *Br. J. Dermatol.* **179**, 506–508 (2018).
70. Morris, D. L. et al. Genome-wide association meta-analysis in Chinese and European individuals identifies ten new loci associated with systemic lupus erythematosus. *Nat. Genet.* **48**, 940–946 (2016).
71. Alarcon-Riquelme, M. E. et al. Genome-wide Association study in an amerindian ancestry population reveals novel systemic lupus erythematosus risk loci and the role of European admixture. *Arthritis Rheumatol.* **68**, 932–943 (2016).
72. Zhang, Y. et al. Genome-wide search followed by replication reveals genetic interaction of CD80 and ALOX5AP associated with systemic lupus erythematosus in Asian populations. *Ann. Rheum. Dis.* **75**, 891–898 (2016).
73. Kunz, M. et al. Genome-wide association study identifies new susceptibility loci for cutaneous lupus erythematosus. *Exp. Dermatol.* **24**, 510–515 (2015).

74. Lei, S. F. & Deng, F. Y. Identification of susceptibility genes for systemic lupus erythematosus with a genome-wide gene-based association study. *Scand. J. Rheumatol.* **43**, 426–428 (2014).
75. Okada, Y. et al. A genome-wide association study identified *AFF1* as a susceptibility locus for systemic lupus erythematosus in Japanese. *PLoS Genet.* **8**, e1002455 (2012).
76. Yang, W. et al. Genome-wide association study in Asian populations identifies variants in *ETSI* and *WDFY4* associated with systemic lupus erythematosus. *PLoS Genet.* **6**, e1000841 (2010).
77. Cervino, A. C., Tsinoremas, N. F. & Hoffman, R. W. A genome-wide study of lupus: preliminary analysis and data release. *Ann. N. Y. Acad. Sci.* **1110**, 131–139 (2007).
78. Langefeld, C. D. et al. Transancestral mapping and genetic load in systemic lupus erythematosus. *Nat. Commun.* **8**, 16021 (2017).
79. The Genomes Project C. A global reference for human genetic variation. *Nature* **526**, 68 (2015).
80. Chang, C. C. et al. Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience* **4**, 7 (2015).
81. Sherry, S. T. et al. dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res.* **29**, 308–311 (2001).
82. Karolchik, D. et al. The UCSC Table Browser data retrieval tool. *Nucleic Acids Res.* **32**, D493–D496 (2004).
83. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114–2120 (2014).
84. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357–359 (2012).
85. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15**, 550 (2014).
86. Pruitt, K. D., Tatusova, T. & Maglott, D. R. NCBI reference sequence (RefSeq): a curated non-redundant reference sequence database of genomes, transcripts and proteins. *Nucleic Acids Res.* **33**, D501–D504 (2005).
87. Cairns, J. et al. CHiCAGO: robust detection of DNA looping interactions in capture Hi-C data. *Genome Biol.* **17**, 127 (2016).
88. Jung, I. et al. A compendium of promoter-centered long-range chromatin interactions in the human genome. *Nat. Genet.* **51**, 1442–1449 (2019).
89. Kerimov, N. et al. eQTL Catalogue: a compendium of uniformly processed human gene expression and splicing QTLs. Preprint at *bioRxiv* <https://doi.org/10.1101/2020.01.29.924266> (2020).
90. Alasoo, K. et al. Shared genetic effects on chromatin and gene expression indicate a role for enhancer priming in immune response. *Nat. Genet.* **50**, 424–431 (2018).
91. Chen, L. et al. Genetic drivers of epigenetic and transcriptional variation in human immune cells. *Cell* **167**, 1398–1414 e1324 (2016).
92. Gutierrez-Arcelus, M. et al. Passive and active DNA methylation and the interplay with genetic variation in gene regulation. *Elife* **2**, e00523 (2013).
93. Lappalainen, T. et al. Transcriptome and genome sequencing uncovers functional variation in humans. *Nature* **501**, 506–511 (2013).
94. Kilpinen, H. et al. Common genetic variation drives molecular heterogeneity in human iPSCs. *Nature* **546**, 370–375 (2017).
95. Nedelec, Y. et al. Genetic ancestry and natural selection drive population differences in immune responses to pathogens. *Cell* **167**, 657–669 e621 (2016).
96. Quach, H. et al. Genetic adaptation and neandertal admixture shaped the immune system of human populations. *Cell* **167**, 643–656 e617 (2016).
97. Schwartzentruber, J. et al. Molecular and functional variation in iPSC-derived sensory neurons. *Nat. Genet.* **50**, 54–61 (2018).
98. Buil, A. et al. Gene-gene and gene-environment interactions detected by transcriptome sequence analysis in twins. *Nat. Genet.* **47**, 88–91 (2015).
99. van de Bunt, M. et al. Transcript expression data from human islets links regulatory signals from genome-wide association studies for type 2 diabetes and glycemic traits to their downstream effectors. *PLoS Genet.* **11**, e1005694 (2015).
100. Schmiedel, B. J. et al. Impact of genetic polymorphisms on human immune. *Cell Gene Expr. Cell* **175**, 1701–1715 e1716 (2018).
101. Jaffe, A. E. et al. Developmental and genetic regulation of the human cortex transcriptome illuminate schizophrenia pathogenesis. *Nat. Neurosci.* **21**, 1117–1125 (2018).
102. Ng, B. et al. An xQTL map integrates the genetic architecture of the human brain's transcriptome and epigenome. *Nat. Neurosci.* **20**, 1418–1426 (2017).
103. Lepik, K. et al. C-reactive protein upregulates the whole blood expression of *CD59*—an integrative analysis. *PLoS Comput. Biol.* **13**, e1005766 (2017).
104. Taylor, D. L. et al. Integrative analysis of gene expression, DNA methylation, physiological traits, and genetic variation in human skeletal muscle. *Proc. Natl Acad. Sci. USA* **116**, 10883–10888 (2019).
105. Fairfax, B. P. et al. Genetics of gene expression in primary immune cells identifies cell type-specific master regulators and roles of HLA alleles. *Nat. Genet.* **44**, 502–510 (2012).
106. Fairfax, B. P. et al. Innate immune activity conditions the effect of regulatory variants upon monocyte gene expression. *Science* **343**, 1246949 (2014).
107. Kasela, S. et al. Pathogenic implications for autoimmune mechanisms derived by comparative eQTL analysis of CD4+ versus CD8+ T cells. *PLoS Genet.* **13**, e1006643 (2017).
108. Naranbhai, V. et al. Genomic modulators of gene expression in human neutrophils. *Nat. Commun.* **6**, 7545 (2015).
109. Degner, J. F. et al. DNase I sensitivity QTLs are a major determinant of human expression variation. *Nature* **482**, 390–394 (2012).
110. Liang, L. et al. A cross-platform analysis of 14,177 expression quantitative trait loci derived from lymphoblastoid cell lines. *Genome Res.* **23**, 716–726 (2013).
111. Raj, T. et al. Polarization of the effects of autoimmune and neurodegenerative risk alleles in leukocytes. *Science* **344**, 519–523 (2014).
112. Lee, M. N. et al. Common genetic variants modulate pathogen-sensing responses in human dendritic cells. *Science* **343**, 1246980 (2014).
113. Chen, E. Y. et al. Enrichr: interactive and collaborative HTML5 gene list enrichment analysis tool. *BMC Bioinforma.* **14**, 128 (2013).
114. Kuleshov, M. V. et al. Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. *Nucleic Acids Res.* **44**, W90–W97 (2016).
115. Barrett, T. et al. NCBI GEO: archive for functional genomics data sets—update. *Nucleic Acids Res.* **41**, D991–D995 (2013).
116. Goodwin, S., McPherson, J. D. & McCombie, W. R. Coming of age: ten years of next-generation sequencing technologies. *Nat. Rev. Genet.* **17**, 333–351 (2016).
117. Bentley, D. R. et al. Accurate whole human genome sequencing using reversible terminator chemistry. *Nature* **456**, 53–59 (2008).
118. Patwardhan, R. P. et al. High-resolution analysis of DNA regulatory elements by synthetic saturation mutagenesis. *Nat. Biotechnol.* **27**, 1173–1175 (2009).
119. Vaughn, S. E. et al. Lupus risk variants in the *PXK* locus alter B-cell receptor internalization. *Front. Genet.* **5**, 450 (2014).
120. Amemiya, H. M., Kundaje, A. & Boyle, A. P. The ENCODE blacklist: identification of problematic regions of the genome. *Sci. Rep.* **9**, 9354 (2019).
121. Lambert, S. A. et al. The human transcription factors. *Cell* **172**, 650–665 (2018).
122. Stormo, G. D. Consensus patterns in DNA. *Methods Enzymol.* **183**, 211–221 (1990).

Acknowledgements

We thank Roger Pique-Regi and Francesca Luca for their consultation and guidance in the development of our MPRA libraries and approach. We thank Kevin Ernst for computational support. We greatly appreciate Aaron Zorn, Raphael Kopan, Marc Rothenberg, and Stephen Wagoner for constructive feedback and guidance. This work was funded by the National Institutes of Health (F32 AI129249, 1K22AI153648-01, K99-HG009920, P30 AR070549, P30 DK078392, R01 AI024717, R01 AI148276, R01 AR073228, R01 DK107502, R01 GM055479, R01 HG010730, R01 NS099068, U01 AI130830, U01 HG008666, U01 AI150748), Cincinnati Children's Hospital Research Foundation (Academic and Research Committee award, Center for Pediatric Genomics pilot funding, Endowed Scholar award), the Veterans Administration (01 BX001834), the Ohio Supercomputing Center, and the State of Ohio.

Author contributions

This manuscript was written by X.L., X.C., M.T.W., and L.C.K. with critical input from C.F., O.D., D.M., S.P., T.H., Y.H., M.P., T.C., E.R.M., J.P.R., C.G.dB., and J.B.H. Experiments were designed by X.L., J.P.R., C.G.dB., M.T.W., and L.C.K. with computational design by X.L., X.C., S.P., J.P.R., C.G.dB., M.T.W., and L.C.K. Experiments were performed by X.L., C.F., O.D., D.M., and Y.H. Analysis was performed by X.L., X.C., and S.P. Data were provided by T.H., M.P., T.C., E.R.M., and M.T.W. Funding was provided by J.P.R., C.G.dB., J.B.H., M.T.W., and L.C.K.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at <https://doi.org/10.1038/s41467-021-21854-5>.

Correspondence and requests for materials should be addressed to M.T.W. or L.C.K.

Peer review information *Nature Communications* thanks James Lee, Kim Simpfordorfer and the other, anonymous, reviewers for their contribution to the peer review of this work.

Reprints and permission information is available at <http://www.nature.com/reprints>

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2021