Genome-wide discovery of SLE genetic risk variant allelic enhancer activity

Xiaoming Lu*[1], Xiaoting Chen*[1], Carmy Forney[1], Omer Donmez[1], Daniel Miller[1], Sreeja Parameswaran[1], Ted Hong[1,2], Yongbo Huang[1], Mario Pujato[3], Tareian Cazares[4], Emily R. Miraldi[3-5], John P. Ray[6], Carl G. de Boer[6], John B. Harley[1,4,5,7,8], Matthew T. Weirauch[#,1,3,5,8,9], Leah C. Kottyan[#,1,4,5,9]

*Contributed equally

[#]Co-corresponding authors: Leah.Kottyan@cchmc.org; Matthew.Weirauch@cchmc.org

[1]Center for Autoimmune Genomics and Etiology, Cincinnati Children's Hospital Medical Center, Cincinnati, Ohio, USA, 45229.

[2]Department of Pharmacology & Systems Physiology, University of Cincinnati, College of Medicine, Cincinnati, Ohio, USA, 45229.

[3]Division of Biomedical Informatics, Cincinnati Children's Hospital Medical Center, Cincinnati, Ohio, USA, 45229.

[4]Division of Immunobiology, Cincinnati Children's Hospital Medical Center, Cincinnati, Ohio, USA, 45229.

[5]Department of Pediatrics, University of Cincinnati, College of Medicine, Cincinnati, Ohio, USA, 45229.

[6]Broad Institute of Massachusetts Institute of Technology (MIT) and Harvard University, Cambridge, Massachusetts, USA, 02142.

[7]US Department of Veterans Affairs Medical Center, Cincinnati, Ohio, USA 45229.

[8]Division of Developmental Biology, Cincinnati Children's Hospital Medical Center, Cincinnati, Ohio, USA, 45229.

[9]Division of Allergy and Immunology, Cincinnati Children's Hospital Medical Center, Cincinnati, Ohio, USA, 45229.

Abstract

Genome-wide association studies of Systemic Lupus Erythematosus (SLE) nominate 3,073 genetic variants at 91 risk loci. To systematically screen these variants for allelic transcriptional enhancer activity, we constructed 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 EBV-infected B cells revealed 482 variants with enhancer activity, with 51 variants showing genotype-dependent (allelic) enhancer activity at 27 risk loci. In-depth analysis of allelic transcription factor (TF) binding at and around these 51 variants identified one class of TFs whose DNA-binding motif tends to be directly altered by the risk variant and a second, larger class of TFs that also bind allelically but do not have their motifs directly altered 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.

Systemic Lupus Erythematosus (SLE) is an autoimmune disease that can affect multiple organs, leading to debilitating inflammation and mortality[1]. 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 SLE[2]. SLE has a relatively high heritability[3]. 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 sizes[4].

Genome-wide association studies (GWASs) have identified 91 genetic risk loci that increase disease risk of SLE in a largely additive fashion[4]. 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 usage[5]. 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 SLE[8]. However, given the large number of candidate variants identified by GWAS, identification of the particular causal variant(s) remains challenging.

SLE is a complex disease that involves multiple cell types[2]. Previous systematic studies demonstrate that SLE risk loci are enriched for B cell specific genes[9] and regulatory regions[10]. 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 production[11]. B cells also present self-antigens to T cells in the development of an autoantigen-focused (i.e., "self") inflammatory response[12]. 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 EBNA2 genomic interactions are concentrated at SLE risk loci in EBV-infected B cells[10]. Accordingly, EBV-infected B cells offer an ideal disease-relevant model for studying the effects of SLE risk variants at many SLE risk loci.

To systematically identify the SLE genetic risk variants that contribute to transcriptional dysregulation in EBV-infected B cells, we designed and applied a massively parallel reporter assay (MPRA)[17,18] (**Figure 1**). MPRA extends standard reporter assays, replacing low-throughput luciferase with high-throughput RNA expression detection. In this study, we used 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. 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 reveal an important role for groups of TFs in the mediation of allelic enhancer activity at plausibly causal SLE risk variants in EBV-infected B cells.

Results

MPRA library design and quality control

We first collected all SLE-associated risk loci reaching genome-wide association significance ($p<5x10^{-8}$) published through March 2018 **(Supplemental Data Set 1)**. Studies of all ancestral

groups were included, and independent risk loci were defined as loci with lead (tag) variant 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 **(Supplemental Data Set 2)**. In total, this procedure identified 3,073 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 study[18] as positive and negative controls to assess the library's performance **(Supplemental Data Set 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 **(Supplemental Data Set 4)**. A total of 12,478 oligos (3,093 variants with 6,239 alleles) were synthesized. For barcoding, the oligos were ligated to a pool of random 20mers. Each uniquely ligated 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 **(Supplemental Figure 1A, Supplemental Data Set 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 measure the ability of alleles to enhance, but not reduce, transcriptional activity. Three aliquots of the library were independently transfected into the EBV-infected B cells 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. This mRNA to DNA ratio measures the enhancing effect of an allele on eGFP expression under the control of a minimal promoter **(Figure 1)**. We observed strong correlation of enhancer activity between experimental replicates (mean pairwise Pearson correlation of 0.99) **(Supplemental Figure 1B, C, D)**. Likewise, calibration variants showed high

accuracy, with 17 of the 20 variants matching the results of a previous study[18] (87.5% sensitivity and 75% specificity), collectively demonstrating a robust experimental system **(Supplemental Data Set 3)**.

Hundreds of SLE risk variants are located in genomic regions with enhancer activity in EBV-infected B cells

Using the SLE MPRA library, we next identified genetic variants capable of driving enhancer activity in EBV-infected B cells GM12878. An SLE risk variant was considered to reside within a region with enhancer activity if an oligo corresponding to any allele had 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 ($p_{FDR}<0.05$) 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 (**Figure 2A, Supplemental Data Set 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 (**Supplemental Data Set 2, Supplemental Data Set 7**) (see Methods). This approach identified 496 genes in total, which are enriched for expected SLE-related pathways such as the interferon pathway, antigen processing

and presentation pathway, and cytokine-related pathways (**Supplemental Figure 2, Supplemental Data Set 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 algorithm[10]. 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 (Chromatin immunoprecipitation with sequencing) for a particular histone mark or TF). For this analysis, we identified, curated, and systematically processed the 571 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, $p_{corrected} < 10^{-13}$) (**Figure 2B, Supplemental Data Set 9**). As expected, we did not identify enrichment for repressive marks such as H3K9me3 or H3K27me3 (**Figure 2B, Supplemental Data Set 9**). Altogether, the genomic features present within enVars confirm that many SLE genetic risk loci likely alter transcriptional regulation in EBV-infected B cells.

We next asked if the genomic binding sites of particular TFs in B cells were enriched within our enVars using RELI and the anti-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[19-21], relative to variants lacking enhancer activity (**Figure 2C**, **Supplemental Data Set 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 risk[10]. We also found significant enrichment for other TFs that have been previously implicated in SLE pathogenesis, such as PAX5[22], MED1[23], IKZF1[24], ELF1[25] and the

EBV-encoded EBNA2 transactivator[10] (**Figure 2C, Supplemental Data Set 9**). As a complementary approach, we next assessed enrichment for TF binding site motifs in the enAllele DNA sequences using HOMER[26] and motifs contained in the Cis-BP database[27] (see Methods). This analysis also revealed enrichment of multiple TF families with known roles in SLE, including ETS, NFκB, and IRF[3] (**Figure 2D, Supplemental Data Set 10**). Many of these same TFs also have enriched ChIP-seq peaks at SLE risk loci[10]. Collectively, these results indicate that particular TFs tend to not only concentrate at SLE risk loci[10], but also concentrate at alleles capable of driving gene expression in EBV-infected B cells.

MPRA identifies 51 SLE risk variants with allelic enhancer activity in EBV-infected 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 pairs 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[18] ; and (3) the oligos had more than a 25% change between any pair of alleles. Using these criteria, we identified 51 SLE risk variants (11% of enVars, 1.7% of all SLE risk variants) as allelic enVars in EBV-infected B cells GM12878 (**Figure 3A, Supplemental Data Set 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 genome[28]. Collectively, these 51 variants represent causal variant candidates for 27 SLE risk loci (30% of all tested loci) (**Supplemental Data Set 12**). For these 27 risk loci, our approach reduced the number of potential causal variants in EBV-infected B cells from an average of 84 variants to an average of two variants per risk locus (**Figure 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 (**Supplemental Data Set 13**). Many of the top results are consistent with our previous study[10], including the enriched presence of general enhancer features such as the H3K27ac histone mark (17 of 51 allelic enVars, 13.6-fold enriched, $p_{corrected} < 10^{-38}$), mediator complex subunit MED1 (17 of 51 allelic enVars, 13.0-fold enriched, $p_{corrected} < 10^{-34}$), and the histone acetyltransferase p300 (16 of 51 allelic enVars, 12.4-fold enriched, $p_{corrected} < 10^{-32}$), along with particular regulatory proteins that participate in “EBV super enhancers”[29] and play key roles in B cells such as ATF7 (15 of 51 allelic enVars, 11.3-fold enriched, $p_{corrected} < 10^{-26}$), Ikaros/IKZF1 (19 of 51 allelic enVars, 9.7-fold enriched, $p_{corrected} < 10^{-25}$), and the NFκB subunit RELA (13 of 51 allelic enVars, 12.4-fold enriched, $p_{corrected} < 10^{-24}$). Also consistent with our previous study[10], we observe strong enrichment for the EBV-encoded EBNA2 protein (7 of 51 allelic enVars, 17.7-fold enriched, $p_{corrected} < 10^{-19}$). 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 pipeline[10] to search for allelic binding events (i.e., allelic imbalance between sequencing read counts) at SLE variants within the 571 GM12878 ChIP-seq datasets. By necessity, this approach is limited to the 13 allelic enVars that are heterozygous in GM12878. In total, this procedure identified four variants with strong allelic imbalance (MARIO ARS value > 0.4) in at least one ChIP-seq dataset (**Supplemental Data Set 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 SLE[30] and rheumatoid arthritis[31] in Europeans, shows 1.7-fold stronger enhancer activity for the reference/non-risk 'C' allele compared to the non-reference/risk 'T' allele (**Figure 4A**). These results are consistent with a previously established eQTL obtained from GTEx[32], which demonstrates higher Complement C4A (*C4A*) expression in EBV-infected B cells for the rs3101018 'C' allele than the 'T' allele (**Figure 4B**). Our MARIO allelic ChIP-seq analysis reveals 11 regulatory proteins that prefer the 'C' allele and 2 that prefer the 'T' allele (**Figure 4C**). Among these, particularly robust signal is obtained for ATF7, with one experiment displaying 77 vs. 18 reads ('C' vs. 'T') and another showing 66 vs. 23 reads ('C' vs. 'T') (**Supplemental Data Set 14**). Moreover, CREB1 and CREM strongly favor the 'C' allele as well (**Supplemental Data Set 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 (**Figure 4D**). Intriguingly, seven additional proteins (FOXK2, PKNOX1, ARID3A, ZNF217, ARNT, MEF2B, and FOXM1) also bind allelically and have known DNA binding motifs[33], but none of them have binding sites altered by the variant. Further, we do not observe allelic chromatin accessibility in available ATAC-seq datasets (9 vs. 7 reads). Together, these results reveal a potentially causative SLE regulatory mechanism involving weaker direct binding of ATF7/CREB/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 ancestries[34] and rheumatoid arthritis in Europeans[35]. rs2069235 displays much stronger enhancer activity for the 'A' (non-reference/risk) allele compared to the 'G' (reference/non-risk) allele (**Figure 4E**), consistent with the established synaptogyrin 1 (*SYNGR1)* eQTL in EBV-infected B cells[32] (**Figure 4F**). Inspection of our allelic ChIP-seq data reveals a larger number of

proteins binding to the 'A' than the 'G' allele (14 vs. 2) (**Figure 4G**). Among these 16 proteins, only IKZF2 and ELF1 have their binding sites directly altered by the variant (**Figure 4H**). Together, these data are consistent with a potentially causative SLE molecular mechanism involving stronger direct binding of IKZF2 to the 'G' non-risk allele and stronger ELF1 binding to the 'A' risk allele, along with indirectly altered binding of multiple TFs to this locus.

Genotype-dependent binding to SLE variants with allelic 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 (**Figure 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 eVars using a proportion test (see Methods) and confirmed that their binding site locations are distributed relative to the variant as expected (**Figure 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 IRF5[36] (**Figure 5C**, **Supplemental Data Set 15**). *Variant adjacent* TFs represent a largely distinct class, but also include several TFs with SLE genetic associations, including NFκB[37], the Ikaros (IKZF) family[38], and HMGA family

members[39] (**Figure 5D, Supplemental Data Set 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.

In summary, through the application of an allelic MPRA library to the EBV-infected B cells GM12878, we identified global transcriptional enhancer activity at 16% of SLE-associated genetic variants (enVars), with particular 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 in enhancer activity in EBV-infected B cells. 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 plausible variants due to linkage disequilibrium. This study is the first direct genome-wide measurement of enhancer activity of 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-infected B cells 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. SLE risk loci that lack allelic enVars in GM12878 might act in other cell types (e.g., T cells) and/or under different conditions (e.g., subsequent to interferon stimulation).

A critical finding of this study is that SLE risk variants with allelic enhancer activity can 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. In general, a given TF can be *variant overlapping* at one locus and *variant adjacent* at another, as exemplified by ATF7 (**Figure 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., IRF5[36], NFκB[37], and ETS1[40]), 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 levels[36], 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 cells[41]. People without *C4A* due to rare, protein-changing mutations are at a greatly increased risk for autoimmune diseases, including type I diabetes and SLE[42]. Further, the risk of developing SLE is 2.62 times higher in subjects with low total *C4A*[41]. Consistent with this observation, the SLE risk allele at this genetic locus identified in our MPRA is associated with lower *C4A* expression (**Figure 4B**). Additionally, the

SLE risk locus encoding *SYNGR1* was recently identified in a high-density genotyping study of subjects with Asian ancestry[34] and also increases disease risk for schizophrenia[43], primary biliary cirrhosis[44], and rheumatoid arthritis[45]. *SYNGR1* 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 *SYNGR1* in other tissues, including developing B cells[46]. eQTL data from EBV-transformed B cell lines (**Figure 4F**) further support our MPRA-based findings of SLE risk genotype-dependent enhancer activity and gene expression at this locus.

GWAS provides important insight into the genetic origins of disease. In conjunction with other genome-scale assays such as ATAC-seq, ChIP-seq, and HiChIP-seq, MPRA reveals 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 a specific cell type. Our integrative analyses reveal specific molecular mechanisms underlying genotype-dependent transcriptional regulation and SLE disease risk. We conclude that MRPA 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 the many complex human diseases.

Online Methods

Variant selection and DNA sequence generation

All SLE-associated genetic risk loci reaching genome-wide significance published through March 2018 were included in this study[24,30,36,40,47-56]. A total of 91 genetic risk loci were used for linkage disequilibrium (LD) expansion (r^2>0.8) based on 1000 Genomes Data[57] in the ancestry(ies) of the initial genetic association using PLINK[58] (**Supplemental Data Set 1**). All expanded variants were

updated to the dbSNP 151 table[59] from the UCSC table browser[60] based on either variant name or genomic location. Unmappable variants were discarded. We also included 20 genetic variants from the Tewhey *et al.*[18] 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 (15bps) were added to each sequence at either end (5'-ACTGGCCGCTTGACG - [170 bp oligo] - CACTGCGGCTCCTGC-3') to make a 200 bp DNA sequence (**Supplemental Data Set 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 (3,093 variants, 6,239 alleles) were obtained from Twist Bioscience.

MPRA experiments

Library Assembly

For assembly of the MPRA library, we followed the procedure described by Tewhey *et al.* [18] 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, 20bps barcodes were added to the synthesized oligos through 24X PCR with 50 μL system, each containing 1.86ng oligo, 25 μL NEBNext® Ultra™ II Q5® Master Mix, 1μM MPRA_v3_F and MPRA_v3_201_R. PCR was performed under the following conditions: 98°C for 2 mins, 12 cycles of (98°C for 10 sec, 60°C for 15 sec, 72°C for 45 sec), 72°C for 5 mins. Amplified product was purified and cloned into SfiI digested pGL4.23ΔxbaΔluc by Gibson assembly at 50°C for 1 hr. The assembled backbone library was purified and then transformed into *Escherichia coli* (*E.coli)* through electroporation (2kV, 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 hrs. 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 50μl 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 mins, 20 cycles of (98°C for 10 sec, 60°C for 15 sec, 72°C for 30 sec), 72°C for 5 mins. 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 hr. 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 to 16 hrs. 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, 0.5μM 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 mins, 20 cycles of (98°C for 10 sec, 60°C for 15 sec, 72°C for 45 sec), 72°C for 5 mins. The amplified product was purified and then inserted into AsiSI digested backbone library through Gibson assembly at 50°C for 1.5 hrs to create the transfection library. The resulting library was re-digested by RecBCD and AsiSI, purified and then transformed into *E.coli* through electroporation (2kV, 200 ohm, 25 μF). Transformed *E.coli* was cultured in 5L of LB Broth buffer supplemented with 100 μg/mL of carbenicillin at 37 °C for 12 to 16 hrs. The plasmid was then extracted using the QIAGEN Endo-free Plasmid Giga Kit.

Oligo and barcode association

The oligo and barcode regions were amplified from the backbone library through 4X PCR with a 100 μL system containing 200 ng plasmid, 50 μL NEBNext® Ultra™ II Q5® Master Mix, 0.5 μM TruSeq_Universal_Adapter_P5 and MPRA_v3_TruSeq_Amp2Sa_F_P7. PCR was performed under the following conditions: 95°C for 20 sec, 6 cycles of (95°C for 20 sec, 62°C for 15 sec, 72°C for 30 sec), 72°C for 2 mins. The product was then purified, and indices were added through a 100μl system containing all purified product, 50μl NEBNext® Ultra™ II Q5® Master Mix, 0.5 μM TruSeq_Universal_Adapter_P5 and index primer. PCR was performed as above, except for only 5 cycles. Samples were purified, molar pooled, and sequenced using 2x125bp 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 μg/mL of streptomycin. Cells were seeded at a density of 5 x 10^5 cells/mL the day before transfection. For triplicate transfections, we collected a total of 5 x 10^7 cells per replicate. Cells were then suspended with 50 μg transfection library plasmid in 400 μL Buffer R. Electroporation was performed with the Neon transfection system in 100μl needles with 3 pulses of 1200V, 20 ms each. After transfection, cells were recovered in 50 mL pre-warmed RPMI medium supplemented only with 10% FBS for 24 hrs. Cells were then collected for preparation of the sequencing library for barcode counting.

Sequencing library for barcode counting

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 μL system with 2.5 μL Turbo DNase and 37.5 μL Turbo DNase Buffer at 37°C for 1 hr. 3.75 μL 10% SDS and 37.5 μL 0.5M EDTA were added to stop DNase with 5 mins of incubation at 75°C. The whole volume was used for eGFP probe hybridization in an 1800 μL system, with 450 μl 20X SSC Buffer, 900 μL

Formamide and 1 μL of each 100 μM Biotin-labeled GFP probe One to Three. The probe hybridization was performed through incubation at 65°C for 2.5 hrs. 200 μL Dynabeads™ MyOne™ Streptavidin C1 was prepared according to the manufacturer's instruction. The beads were suspended in 250 μL 20X SSC Buffer and incubated with the above probe hybridization reaction at room temperature for 15 mins. 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 μL ddH2O, heating at 70°C for 2 mins and collecting on a magnet, then adding another 12.5 μL ddH2O, heating at 80°C for 2 mins and collecting on a magnet. All collected elution was performed with another DNase treatment in a 30 μL system containing 0.5 μL Turbo DNase and 3 μL Turbo DNase Buffer at 37°C for 1 hr. 0.5 μL 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.*[18]. A total of two PCR were needed for building the sequencing library. The first PCR was performed with TruSeq_Universal_Adapter_P5 and MPRA_V3_Illumina_GFP_F. The second PCR was performed with TruSeq_Universal_Adapter_P5 and index primer. Samples were purified, molar pooled, and sequenced using 1x75bp on Illumina NextSeq 500.

All primers used in this study are provided in **Supplemental Table 1.**

MPRA data analysis

Oligo and barcode association

Paired-end, 125 bp reads were first quality filtered using Trimmomatic-0.38[61] (flags: PE -phred33, LEADING:25, TRAILING:25, MINLEN:80). Read 1 was then separated into the 20bp 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[62] (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 bp reads were first quality filtered using Trimmomatic-0.38[61] (flags: PE -phred33, LEADING:3, TRAILING:3, MINLEN:70). Each read was then separated into the 20bp 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[62] (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.*[18] with minor modifications. In brief, oligos (alleles) with 30 or more unique barcodes from the plasmid control were included for analysis. Barcode count totals for each oligo, including all SLE variants and the 20 control variants, were passed into DESeq2[63] (R version 3.5.3[64]) to estimate the fold change and significance between plasmid controls and the triplicate experiments. A Benjamini-Hochberg FDR adjusted p-value of less than 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.*[18] 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[18]. P-values were adjusted with the Benjamini-Hochberg FDR-based procedure. A corrected p-value of less than 0.05 was required for significance. Only significant alleles with 25%-fold changes or greater were identified as allelic enVars. We have created an R package (mpraprofiler) for performing this analysis, which is available on the Weirauch lab GitHub page (https://github.com/WeirauchLab/).

Downstream MPRA analysis

Gene annotation

We annotated each SLE genetic variant with its nearest gene using the NCBI RefSeq table[65] downloaded from the UCSC table browser[60]. enVars were annotated using a combination of DNA looping interactions (GM12878 Capture Hi-C data[66,67]) and eQTL data obtained from EBV-infected B cell lines (GTEx Analysis V7 (dbGaP Accession phs000424.v7.p2)[32] and other individual studies[68-71]). For all variants, the target genes were annotated (**Supplemental Data Set 2**) using the union of promoter interacting genes and eQTL genes, when available. Otherwise, target genes were annotated as the nearest gene. Allelic enVars 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.

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[26] 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 CisBP database[27] and a log base 2 log likelihood scoring system.

GO enrichment analysis

Enrichr[72,73] 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 (**Supplemental Data Set 8, Supplemental Figure 2)**.

Identification and processing of publicly available GM12878 ChIP-seq data

571 ChIP-seq datasets were obtained from the Gene Expression Omnibus (GEO)[74] using custom scripts that searched for ChIP-seq experiments performed in the GM12878 cell line. 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)[75]. If FastQC detects adapter sequences, the pipeline runs the FastQ files through Trim Galore (v0.4.2)[76], a wrapper script that runs cutadapt (v1.9.1)[77] 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)[62]. The aligned reads (in .BAM format) are then sorted using samtools (v1.8.0)[78] and duplicate reads are removed using picard (v1.89)[79]. Finally, peaks are called using MACS2 (v2.1.2) (flags: callpeak -g hs -q 0.01 -f BAM)[80]. ENCODE blacklist regions[81] 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[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 less than 0.05 and more than 10%- fold change in the DESeq2 result). A distribution of expected overlap values is then created from 2,000 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 data set.

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 browser[60].

Identification of allelic ChIP-seq reads using MARIO

To identify possible mechanisms underlying our allelic enVars, we applied our MARIO[10] method to the GM12878 ChIP-seq dataset collection described above. In brief, MARIO identifies common

genetic variants that are (1) heterozygous in the assayed cell line (here, GM12878) 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 values >0.4 were considered allelic, and any TF with inconsistent allelic imbalance across ChIP-seq datasets at a given variant site was removed, following our previous study[10].

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 sites[82]). 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 frequency matrix (PFM) models contained in the Cis-BP database[27] and the log-likelihood PFM scoring system[83]. 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 PFM 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 PFM score of 70% or higher, and the other allele has a score of less than 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 **Supplemental Data Set 15**.

We used a similar procedure to identify particular TFs that tend to be *variant adjacent* TFs at SLE allelic enVars. We performed HOMER motif enrichment analysis using the full 170bp allelic enVar DNA sequences as input. The standard HOMER null model (scrambled input sequence, maintaining dinucleotide frequencies) was 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 **Supplemental Data Set 15**.

Data availability

MPRA data are available in the Gene Expression Omnibus (GEO) database under accession number GSE143792. This dataset will remain private until the article is "in press".

Full datasets and processed results are provided in the Supplementary Material. Investigators requesting access to the supplementary materials are requested to contact the corresponding authors directly.

Code availability

Source code, with full documentation and examples, are freely available under the GNU General Public License on the Weirauch Lab GitHub page:

https://github.com/WeirauchLab

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 Waggoner for constructive feedback and guidance.

This work was funded by the National Institutes of Health (P30 AR070549, R01 NS099068, U01 AI130830, R01 AR073228, R01 DK107502, R01 AI024717, R01 GM055479, F32 AI129249, K99-HG009920, U01 HG008666, R01 AI148276, P30 DK078392*),* Cincinnati Children's Hospital Research Foundation (Academic and Research Committee award, Center for Pediatric Genomics pilot funding, Endowed Scholar award), the Veterans Administration (I01 BX001834), the Ohio Supercomputing Center, and the State of Ohio.

Author Contributions

This manuscript was written by XL, XC, MTW, and LCK with critical input from CF, OD, DM, SP, TH, YH, MP, TC, ERM, JPR, CGdB, and JBH. Experiments were designed by XL, JPR, CGdB, MTW, and LCK with computational design by XL, XC, SP, JPR, CGdB, MTW, and LCK. Experiments were performed by XL, CF, OD, DM, and YH. Analysis was performed by XL, XC,

and SP. Data were provided by TH, MP, TC, ERM, and MTW. Funding was provided by JPR, CGdB, JBH, MTW, and LCK.

Competing Interests Statement. None of the authors have conflicts to disclose.

References

1. Carter, E.E., Barr, S.G. & Clarke, A.E. The global burden of SLE: prevalence, health disparities and socioeconomic impact. *Nat Rev Rheumatol* **12**, 605-20 (2016).
2. Tsokos, G.C. Systemic Lupus Erythematosus. *New England Journal of Medicine* **365**, 2110-2121 (2011).
3. Deng, Y. & Tsao, B.P. Genetic susceptibility to systemic lupus erythematosus in the genomic era. *Nat Rev Rheumatol* **6**, 683-92 (2010).
4. Visscher, P.M. *et al.* 10 Years of GWAS Discovery: Biology, Function, and Translation. *Am J Hum Genet* **101**, 5-22 (2017).
5. 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).
6. 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).
7. Farh, K.K. *et al.* Genetic and epigenetic fine mapping of causal autoimmune disease variants. *Nature* **518**, 337-43 (2015).
8. 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).
9. 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).
10. Harley, J.B. *et al.* Transcription factors operate across disease loci, with EBNA2 implicated in autoimmunity. *Nat Genet* **50**, 699-707 (2018).
11. 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).
12. 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**, 85-9 (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**, 3019-26 (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. Patwardhan, R.P. *et al.* Massively parallel functional dissection of mammalian enhancers in vivo. *Nat Biotechnol* **30**, 265-70 (2012).
18. Tewhey, R. *et al.* Direct Identification of Hundreds of Expression-Modulating Variants using a Multiplexed Reporter Assay. *Cell* **165**, 1519-1529 (2016).
19. Schebesta, A. *et al.* Transcription Factor Pax5 Activates the Chromatin of Key Genes Involved in B Cell Signaling, Adhesion, Migration, and Immune Function. *Immunity* **27**, 49-63 (2007).
20. Bartels, M. *et al.* Peptide-mediated disruption of NFkappaB/NRF interaction inhibits IL-8 gene activation by IL-1 or Helicobacter pylori. *J Immunol* **179**, 7605-13 (2007).

21. Mancini, M. *et al.* Rel-Dependent Immune and Central Nervous System Mechanisms Control Viral Replication and Inflammation during Mouse Herpes Simplex Encephalitis. *J Immunol* **202**, 1479-1493 (2019).
22. 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**, 265-79 (2015).
23. Armstrong, D.L. *et al.* GWAS identifies novel SLE susceptibility genes and explains the association of the HLA region. *Genes Immun* **15**, 347-54 (2014).
24. 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**, 1234-7 (2009).
25. Yang, J. *et al.* ELF1 is associated with systemic lupus erythematosus in Asian populations. *Hum Mol Genet* **20**, 601-7 (2011).
26. 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**, 576-89 (2010).
27. Weirauch, M.T. *et al.* Determination and inference of eukaryotic transcription factor sequence specificity. *Cell* **158**, 1431-1443 (2014).
28. Stergachis, A.B. *et al.* Exonic transcription factor binding directs codon choice and affects protein evolution. *Science* **342**, 1367-72 (2013).
29. Zhao, B. *et al.* Epstein-Barr virus exploits intrinsic B-lymphocyte transcription programs to achieve immortal cell growth. *Proc Natl Acad Sci U S A* **108**, 14902-7 (2011).
30. International Consortium for Systemic Lupus Erythematosus, G. *et al.* Genome-wide association scan in women with systemic lupus erythematosus identifies susceptibility variants in ITGAM, PXK, KIAA1542 and other loci. *Nat Genet* **40**, 204-10 (2008).
31. 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 Suppl 7**, S129 (2009).
32. Consortium, G.T. The Genotype-Tissue Expression (GTEx) project. *Nat Genet* **45**, 580-5 (2013).
33. Lambert, S.A. *et al.* Similarity regression predicts evolution of transcription factor sequence specificity. *Nat Genet* **51**, 981-989 (2019).
34. Sun, C. *et al.* High-density genotyping of immune-related loci identifies new SLE risk variants in individuals with Asian ancestry. *Nat Genet* **48**, 323-30 (2016).
35. Okada, Y. *et al.* Genetics of rheumatoid arthritis contributes to biology and drug discovery. *Nature* **506**, 376-81 (2014).
36. 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**, 582-96 (2015).
37. Li, Y. *et al.* Genetic Variants of IkappaB Kinase beta (IKBKB) and Polymerase beta (POLB) Were Not Associated with Systemic Lupus Erythematosus Risk in a Chinese Han Population. *PLoS One* **10**, e0132556 (2015).
38. 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).
39. 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).
40. Lu, X. *et al.* Lupus Risk Variant Increases pSTAT1 Binding and Decreases ETS1 Expression. *Am J Hum Genet* **96**, 731-9 (2015).
41. 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).

42. 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).
43. Iatropoulos, P. *et al.* Association study and mutational screening of SYNGR1 as a candidate susceptibility gene for schizophrenia. *Psychiatr Genet* **19**, 237-43 (2009).
44. 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).
45. Liu, J.Z. *et al.* Dense fine-mapping study identifies new susceptibility loci for primary biliary cirrhosis. *Nat Genet* **44**, 1137-41 (2012).
46. Wu, C. *et al.* BioGPS: an extensible and customizable portal for querying and organizing gene annotation resources. *Genome Biol* **10**, R130 (2009).
47. 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).
48. 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).
49. 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**, 932-43 (2016).
50. 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**, 891-8 (2016).
51. Kunz, M. *et al.* Genome-wide association study identifies new susceptibility loci for cutaneous lupus erythematosus. *Exp Dermatol* **24**, 510-5 (2015).
52. 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**, 426-8 (2014).
53. Okada, Y. *et al.* A genome-wide association study identified AFF1 as a susceptibility locus for systemic lupus eyrthematosus in Japanese. *PLoS Genet* **8**, e1002455 (2012).
54. Yang, W. *et al.* Genome-wide association study in Asian populations identifies variants in ETS1 and WDFY4 associated with systemic lupus erythematosus. *PLoS Genet* **6**, e1000841 (2010).
55. 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**, 131-9 (2007).
56. Langefeld, C.D. *et al.* Transancestral mapping and genetic load in systemic lupus erythematosus. *Nat Commun* **8**, 16021 (2017).
57. The Genomes Project, C. *et al.* A global reference for human genetic variation. *Nature* **526**, 68 (2015).
58. Chang, C.C. *et al.* Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience* **4**, 7 (2015).
59. Sherry, S.T. *et al.* dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res* **29**, 308-11 (2001).
60. Karolchik, D. *et al.* The UCSC Table Browser data retrieval tool. *Nucleic Acids Res* **32**, D493-6 (2004).
61. Bolger, A.M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114-20 (2014).
62. Langmead, B. & Salzberg, S.L. Fast gapped-read alignment with Bowtie 2. *Nat Methods* **9**, 357-9 (2012).
63. 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).
64. Team, R.C. R: A Language and Environment for Statistical Computing. (2019).

65. Pruitt, K.D., Tatusova, T. & Maglott, D.R. NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res* **33**, D501-4 (2005).
66. Cairns, J. *et al.* CHiCAGO: robust detection of DNA looping interactions in Capture Hi-C data. *Genome Biol* **17**, 127 (2016).
67. Jung, I. *et al.* A compendium of promoter-centered long-range chromatin interactions in the human genome. *Nat Genet* **51**, 1442-1449 (2019).
68. Degner, J.F. *et al.* DNase I sensitivity QTLs are a major determinant of human expression variation. *Nature* **482**, 390-4 (2012).
69. Liang, L. *et al.* A cross-platform analysis of 14,177 expression quantitative trait loci derived from lymphoblastoid cell lines. *Genome Res* **23**, 716-26 (2013).
70. Raj, T. *et al.* Polarization of the effects of autoimmune and neurodegenerative risk alleles in leukocytes. *Science* **344**, 519-23 (2014).
71. Lee, M.N. *et al.* Common genetic variants modulate pathogen-sensing responses in human dendritic cells. *Science* **343**, 1246980 (2014).
72. Chen, E.Y. *et al.* Enrichr: interactive and collaborative HTML5 gene list enrichment analysis tool. *BMC Bioinformatics* **14**, 128 (2013).
73. Kuleshov, M.V. *et al.* Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. *Nucleic Acids Res* **44**, W90-7 (2016).
74. Barrett, T. *et al.* NCBI GEO: archive for functional genomics data sets--update. *Nucleic Acids Res* **41**, D991-5 (2013).
75. Kalita, C.A. *et al.* QuASAR-MPRA: accurate allele-specific analysis for massively parallel reporter assays. *Bioinformatics* **34**, 787-794 (2018).
76. Goodwin, S., McPherson, J.D. & McCombie, W.R. Coming of age: ten years of next-generation sequencing technologies. *Nat Rev Genet* **17**, 333-51 (2016).
77. Bentley, D.R. *et al.* Accurate whole human genome sequencing using reversible terminator chemistry. *Nature* **456**, 53-9 (2008).
78. Patwardhan, R.P. *et al.* High-resolution analysis of DNA regulatory elements by synthetic saturation mutagenesis. *Nat Biotechnol* **27**, 1173-5 (2009).
79. Vaughn, S.E. *et al.* Lupus risk variants in the PXK locus alter B-cell receptor internalization. *Front Genet* **5**, 450 (2014).
80. 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**, 582-96 (2015).
81. Amemiya, H.M., Kundaje, A. & Boyle, A.P. The ENCODE Blacklist: Identification of Problematic Regions of the Genome. *Sci Rep* **9**, 9354 (2019).
82. Lambert, S.A. *et al.* The Human Transcription Factors. *Cell* **172**, 650-665 (2018).
83. Stormo, G.D. Consensus patterns in DNA. *Methods Enzymol* **183**, 211-21 (1990).

Figure Legends

Figure 1. Massively Parallel Reporter Assay Workflow. See text for description.

Figure 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. Enhancer alleles (enAlleles) were identified as those alleles with significant activity relative to control ($p_{adj} < 0.05$) and at least a 50% increase in activity (see Methods). b. Enrichment of histone marks in GM12878 cells at enVars compared to non-enVars. p-values were estimated using RELI (see Methods). Full RELI results are provided in **Supplemental Data Set 9**. c. Enrichment of regulatory protein and transcription factor (TF) binding at enVars compared to non-enVars. The top 15 TFs (based on RELI p-values) that overlap at least 10% of enVars are shown. Full results are provided in **Supplemental Data Set 9**. d. TF binding site motif enrichment for enVars compared to non-enVars. p-values were obtained from 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 **Supplemental Data Set 10**.

Figure 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 (see Methods). MPRA enhancer activity (X-axis) is presented as the maximum normalized fold-change of MPRA activity for any allele of the variant. Allelic enVars were defined as variants with a significant difference in MPRA activity ($p_{adj} < 0.05$) between any pair of alleles and at least a 25% change in activity difference (see Methods). 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 grey.

Figure 4. Lupus risk allele-dependent gene regulatory mechanisms at the *C4A* and *SYNGR1* genomic loci. a. and e. Normalized MPRA enhancer activity of each experimental replicate for rs3101018 and rs26069235. b. and f. Expression trait quantitative loci (eQTLs) revealing genotype-dependent expression of *C4A* and *SNYGR1* for rs3101018 and rs26069235 in EBV-infected B cell lines (GTEx). c. and g. Genotype-dependent activity of transcription factors, transcriptional regulators, and histone marks in GM12878 cells for rs3101018 and rs26069235. Results with MARIO ARS values >0.4, at least a 1.5-fold imbalance between alleles, and consistent allelic imbalance across TF 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). *Variant overlapping* TFs are indicated in black. *Variant adjacent* TFs are shown in green (see definition in **Figure 5a**). d. and h. DNA binding motif logos are shown for the ATF/CREB/CREM family, IKZF2, and ELF1 in the context of the DNA sequence surrounding rs3101018 and rs2069235, respectively. Tall nucleotides above the X-axis indicate preferred DNA bases. Bases below the X-axis are disfavored.

Figure 5. Identification of *variant overlapping* and *variant adjacent* TFs. a. Model of *variant overlapping* and *variant adjacent* transcription factors (TFs). *Variant overlapping* TFs allelically bind on top of variants, while *variant adjacent* TFs allelically bind near variants. b. TF binding site location distribution for *variant overlapping* and *variant adjacent* 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 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 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.

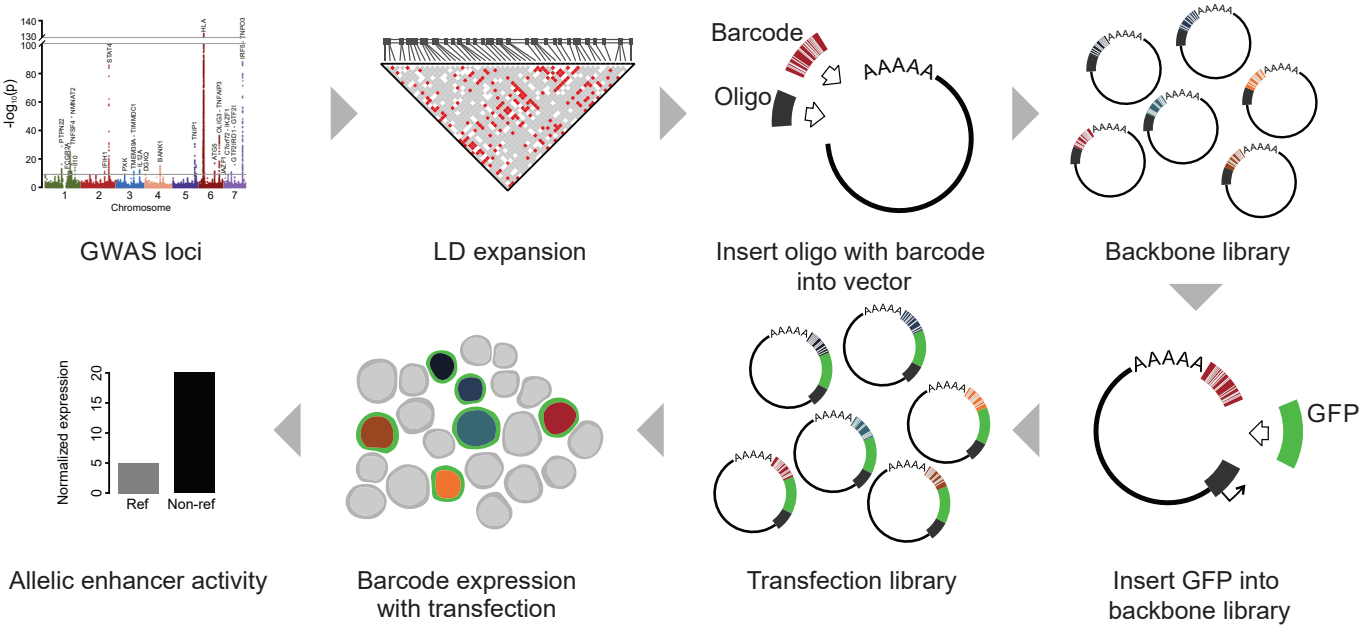
-log10(p)
Chromosome
GWAS loci
LD expansion
Barcode
Oligo
AAAAA
Insert oligo with barcode into vector
Backbone library
GFP
Insert GFP into backbone library
Transfection library
Barcode expression with transfection
Normalized expression
Ref
Non-ref
Allelic enhancer activity

Figure 1

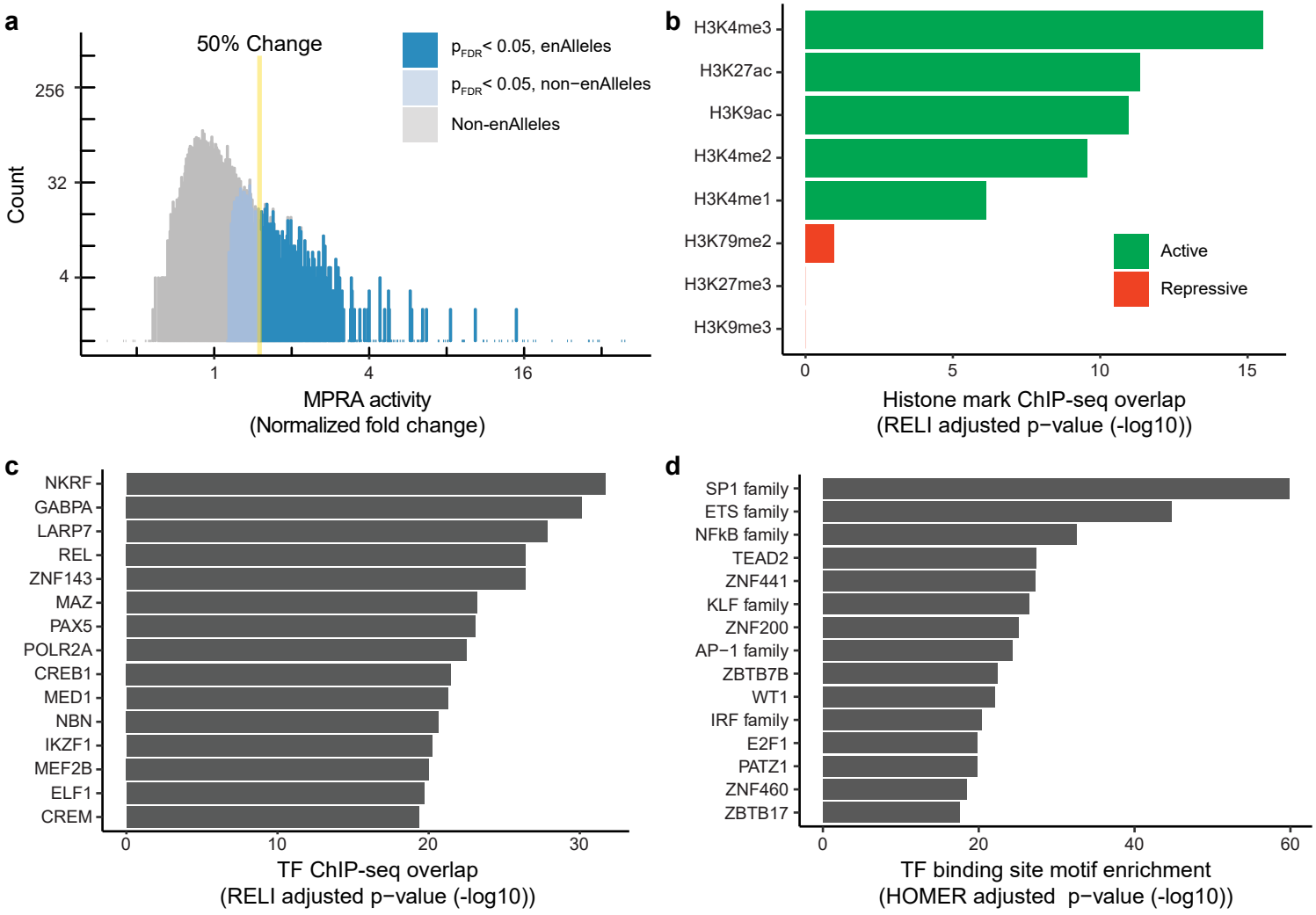
a
50% Change
p_{FDR}< 0.05, enAlleles
p_{FDR}< 0.05, non−enAlleles
Non-enAlleles
Count
256
32
4
1
4
16
MPRA activity
(Normalized fold change)
b
H3K4me3
H3K27ac
H3K9ac
H3K4me2
H3K4me1
H3K79me2
H3K27me3
H3K9me3
Active
Repressive
0
5
10
15
Histone mark ChIP-seq overlap
(RELI adjusted p−value (-log10))
c
NKRF
GABPA
LARP7
REL
ZNF143
MAZ
PAX5
POLR2A
CREB1
MED1
NBN
IKZF1
MEF2B
ELF1
CREM
0
10
20
30
TF ChIP-seq overlap
(RELI adjusted p−value (-log10))
d
SP1 family
ETS family
NFkB family
TEAD2
ZNF441
KLF family
ZNF200
AP−1 family
ZBTB7B
WT1
IRF family
E2F1
PATZ1
ZNF460
ZBTB17
0
20
40
60
TF binding site motif enrichment
(HOMER adjusted p−value (-log10))

Figure 2

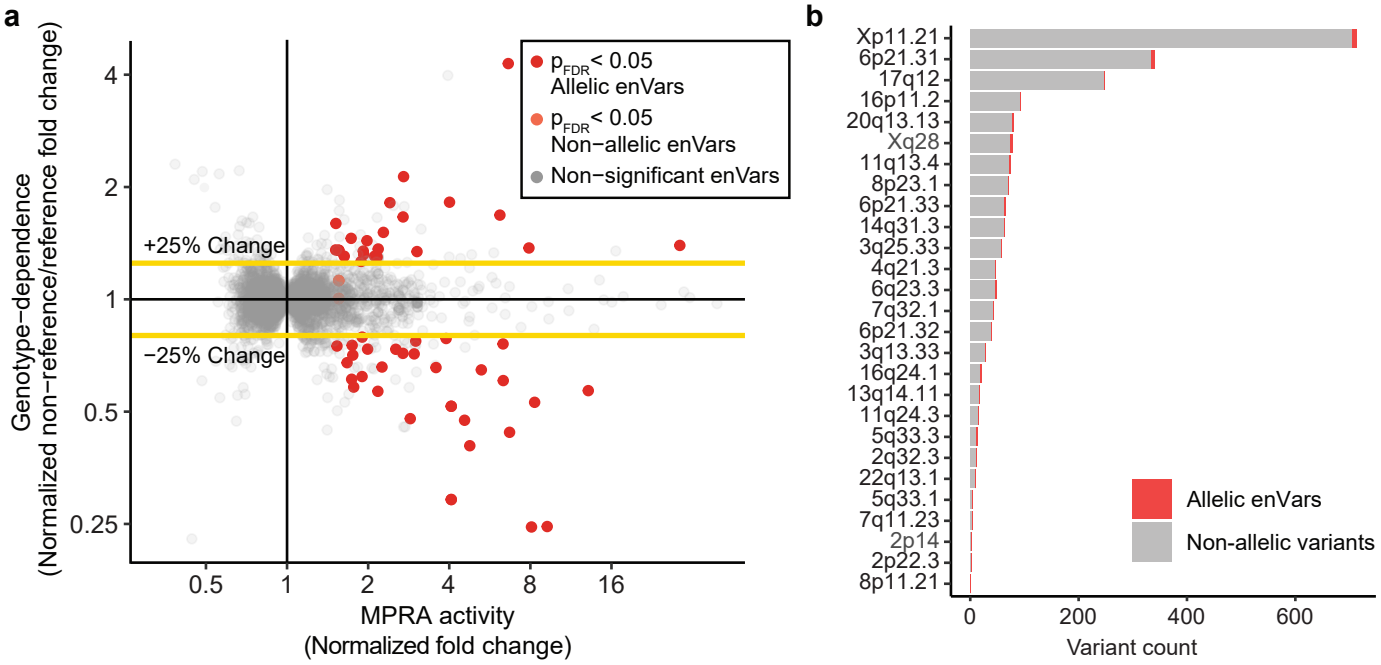
a
Genotype−dependence
(Normalized non−reference/reference fold change)
4
2
1
0.5
0.25
+25% Change
−25% Change
p_{FDR}< 0.05
Allelic enVars
p_{FDR}< 0.05
Non−allelic enVars
Non−significant enVars
0.5
1
2
4
8
16
MPRA activity
(Normalized fold change)
b
Xp11.21
6p21.31
17q12
16p11.2
20q13.13
Xq28
11q13.4
8p23.1
6p21.33
14q31.3
3q25.33
4q21.3
6q23.3
7q32.1
6p21.32
3q13.33
16q24.1
13q14.11
11q24.3
5q33.3
2q32.3
22q13.1
5q33.1
7q11.23
2p14
2p22.3
8p11.21
Allelic enVars
Non-allelic variants
0
200
400
600
Variant count

Figure 3

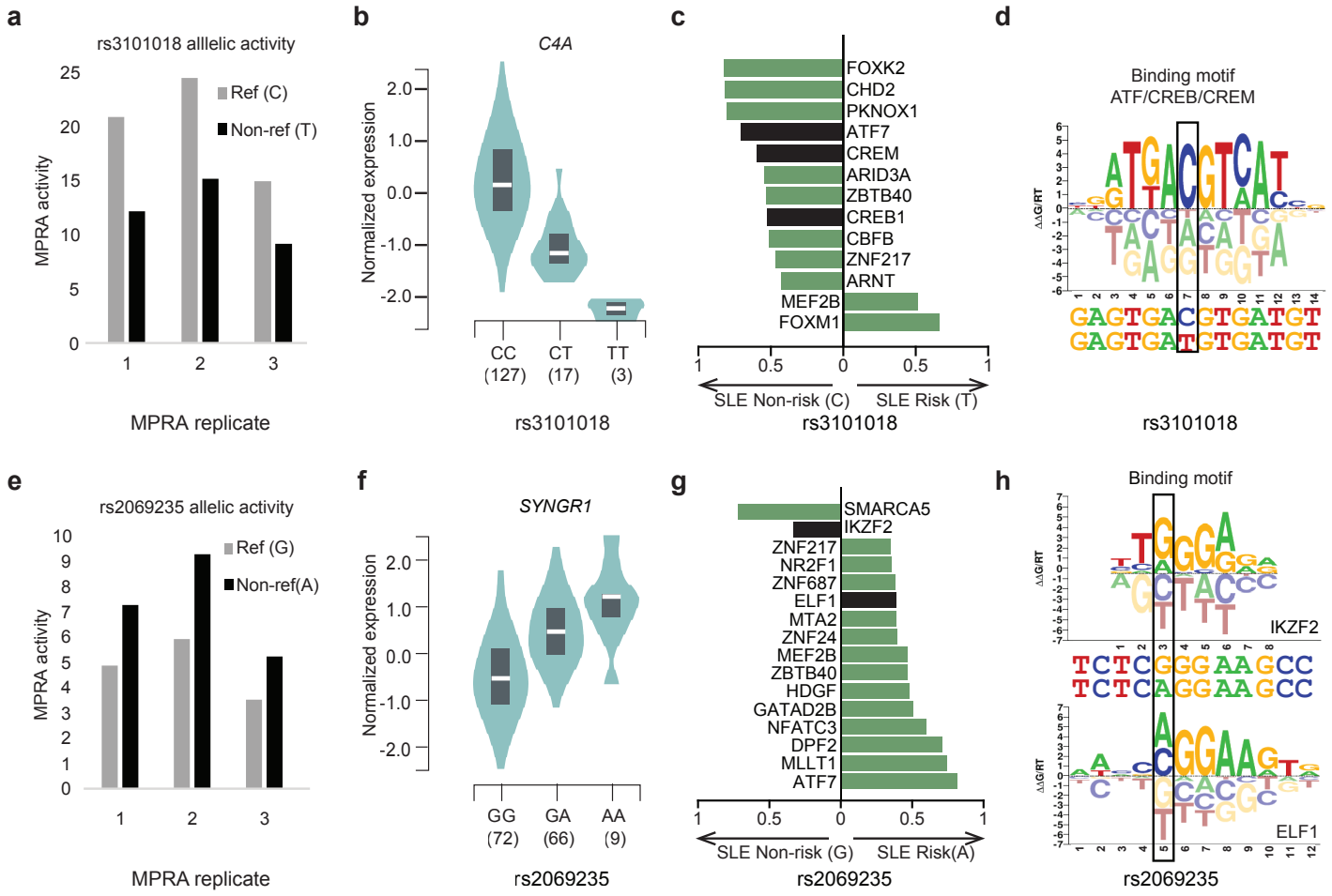
a
rs3101018 alllelic activity
Ref (C)
Non-ref (T)
MPRA activity
MPRA replicate
b
C4A
Normalized expression
CC (127)
CT (17)
TT (3)
rs3101018
c
FOXK2
CHD2
PKNOX1
ATF7
CREM
ARID3A
ZBTB40
CREB1
CBFB
ZNF217
ARNT
MEF2B
FOXM1
SLE Non-risk (C)
SLE Risk (T)
rs3101018
d
Binding motif
ATF/CREB/CREM
rs3101018
e
rs2069235 allelic activity
Ref (G)
Non-ref(A)
MPRA activity
MPRA replicate
f
SYNGR1
Normalized expression
GG (72)
GA (66)
AA (9)
rs2069235
g
SMARCA5
IKZF2
ZNF217
NR2F1
ZNF687
ELF1
MTA2
ZNF24
MEF2B
ZBTB40
HDGF
GATAD2B
NFATC3
DPF2
MLLT1
ATF7
SLE Non-risk (G)
SLE Risk(A)
rs2069235
h
Binding motif
IKZF2
ELF1
rs2069235

Figure 4

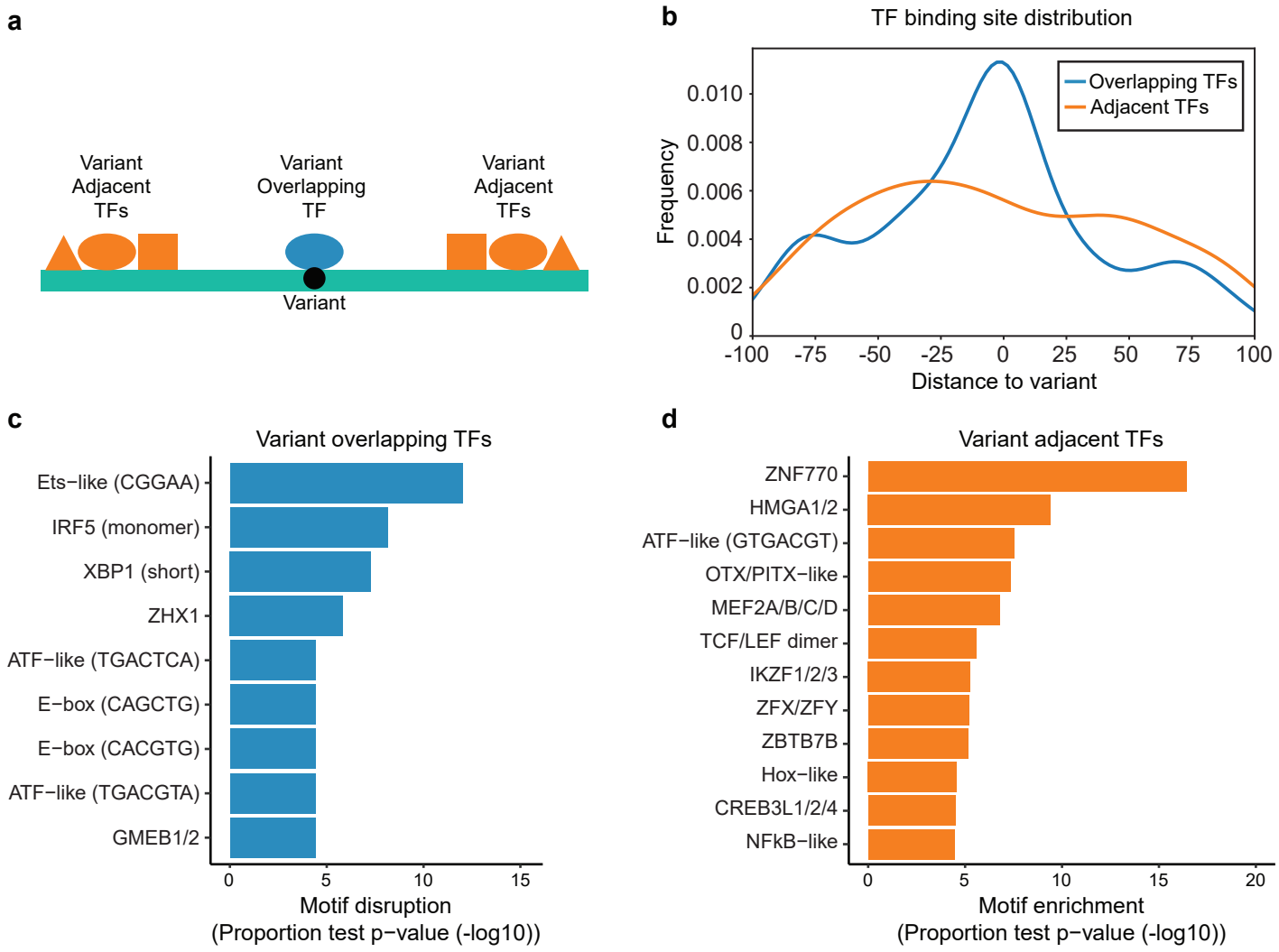
a
Variant Adjacent TFs
Variant Overlapping TF
Variant Adjacent TFs
Variant
b
TF binding site distribution
Overlapping TFs
Adjacent TFs
Frequency
0.010
0.008
0.006
0.004
0.002
0
-100
-75
-50
-25
0
25
50
75
100
Distance to variant
c
Variant overlapping TFs
Ets−like (CGGAA)
IRF5 (monomer)
XBP1 (short)
ZHX1
ATF−like (TGACTCA)
E−box (CAGCTG)
E−box (CACGTG)
ATF−like (TGACGTA)
GMEB1/2
0
5
10
15
Motif disruption
(Proportion test p−value (-log10))
d
Variant adjacent TFs
ZNF770
HMGA1/2
ATF−like (GTGACGT)
OTX/PITX−like
MEF2A/B/C/D
TCF/LEF dimer
IKZF1/2/3
ZFX/ZFY
ZBTB7B
Hox−like
CREB3L1/2/4
NFkB−like
0
5
10
15
20
Motif enrichment
(Proportion test p−value (-log10))

Figure 5