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.

24 [9]Division of Allergy and Immunology, Cincinnati Children's Hospital Medical Center, Cincinnati,
25 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 the Epstein-Barr virus-transformed B cell line GM12878 revealed 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, larger 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.

40 Systemic Lupus Erythematosus (SLE) is an autoimmune disease that can affect multiple organs,
41 leading to debilitating inflammation and mortality[1]. Up to 150 cases are found per 100,000
42 individuals, and the limited treatment options contribute to considerable economic and social
43 burden[1,2]. Epidemiological studies have established a role for both genetic and environmental
44 factors in the development of SLE[2]. SLE has a relatively high heritability[3]. The vast majority of
45 patients do not have a single disease-causing mutation (such as mutations in complement protein
46 1q); instead, genetic risk is accumulated additively through many genetic risk loci with modest
47 effect sizes[4].

48 Genome-wide association studies (GWASs) have identified 91 genetic risk loci that increase
49 disease risk of SLE in a largely additive fashion[4]. Each SLE risk locus is a segment of the genome
50 containing a polymorphic "tag" variant (i.e. the variant with the most significant GWAS p-value)
51 and the genetic variants in linkage disequilibrium with the tag variant. The majority (68%) of the
52 established SLE risk loci do not contain a disease associated coding variant that changes amino
53 acid usage[5]. Instead, variants at these loci are found in non-coding regions of the genome such
54 as introns, promoters, enhancers, and other intergenic areas. Enrichment of these variants in
55 enhancers and at transcription factor (TF) binding sites[6,7] implies that transcriptional perturbation
56 may be a key to the development of SLE[8]. However, given the large number of candidate variants
57 identified by GWASs, identification of the particular causal variant(s) remains challenging.

58 SLE is a complex disease that involves multiple cell types[2]. Previous systematic studies
59 demonstrate that SLE risk loci are enriched for B cell specific genes[9] and regulatory regions[10].
60 Established biological mechanisms further highlight a key role for B cells in SLE - as the
61 autoantibody-secreting cell type, B cells are critical to the pathoetiology of SLE, a disease
62 characterized by autoantibody production[11]. B cells also present self-antigens to T cells in the
63 development of an autoantigen-focused (i.e. "self") inflammatory response[12]. Meanwhile, Epstein-
64 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 lines[10]. *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 many SLE risk loci.

To systematically identify the SLE genetic risk variants that contribute to transcriptional dysregulation in the EBV-transformed B cell line GM12878, we designed and applied a massively parallel reporter assay (MPRA)[18-27] (**Figure 1, Supplemental Figure 1**). MPRA extends standard reporter assays, replacing low-throughput luciferase with high-throughput mRNA 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. 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.

90

91 **Results**

92 **MPRA library design and quality control**

93 We first collected all SLE-associated risk loci reaching genome-wide association significance
94 ($p<5x10^{-8}$) published through March 2018 **(Supplemental Data Set 1)**. Studies of all ancestral
95 groups were included, and independent risk loci were defined as loci with lead (tag) variant at
96 $r^2<0.2$. For each of these 91 risk loci, we performed linkage disequilibrium (LD) expansion ($r^2>0.8$)
97 in each ancestry of the initial genetic association(s), to include all possible disease-relevant
98 variants **(Supplemental Data Set 2)**. In total, this procedure identified 3,073 genetic variants. All
99 established alleles of these variants were included, with 36 variants having three or more alleles.
100 We also included 20 additional genetic variants from a previously published study[19] as positive
101 and negative controls to assess the library's performance **(Supplemental Data Set 3)**.

102 For each variant, we generated a pair of 170 base pair (bp) DNA oligonucleotides (subsequently
103 referred to as "oligos") for each allele, with the variant located in the center and identical flanking
104 genomic sequence across the alleles **(Supplemental Data Set 4)**. A total of 12,478 oligos (3,093
105 variants with 6,239 alleles) were synthesized. For barcoding, a pool of random 20mers were
106 added to the oligo through PCR. Each unique barcode was matched with perfectly synthesized
107 oligos. The number of unique barcodes per oligo had an approximately normal distribution with a
108 median of 729 barcodes per oligo **(Supplemental Figure 2A, Supplemental Data Set 5)**. Only
109 oligos with at least 30 unique barcodes were used for downstream analyses. A fragment
110 containing a minimal promoter and an eGFP gene was inserted between the oligo and barcode
111 to generate the MPRA transfection library. We note that the use of a minimal promoter allows us
112 to effectively measure the ability of alleles to enhance, but not reduce, transcriptional activity.
113 Three aliquots of the library were independently transfected into the EBV-transformed B cell line

114 GM12878. We then used nucleic acid capture to enrich for eGFP mRNA and sequenced the
115 barcode region. The normalized barcode ratio between the eGFP mRNA and the plasmid DNA
116 was used to quantify the amount of enhancer activity driven by each oligo (see **Supplemental**
117 **Note 1, Supplemental Figure 5F, G**). This mRNA to DNA ratio measures the enhancing effect
118 of an allele on eGFP expression under the control of a minimal promoter **(Figure 1, Supplemental**
119 **Figure 1)**. We observed strong correlation of enhancer activity between experimental replicates
120 (mean pairwise Pearson correlation of 0.99) **(Supplemental Figure 2B, C, D)**. Likewise,
121 calibration variants showed high accuracy, with 17 of the 20 variants matching the results of a
122 previous study[19] (87.5% sensitivity and 75% specificity), collectively demonstrating a robust
123 experimental system **(Supplemental Data Set 3)**.

124

125 **Hundreds of SLE risk variants are located in genomic regions with enhancer activity in**
126 **EBV-transformed B cells**

127 Using the SLE MPRA library, we next identified genetic variants capable of driving enhancer
128 activity in the EBV-transformed B cell line GM12878. An SLE risk variant was considered a
129 candidate for enhancer activity if an oligo corresponding to any allele had significantly increased
130 transcriptional regulatory activity compared to controls (see Methods). Not all statistically
131 significant changes in transcriptional activity are necessarily biologically relevant – a highly
132 consistent, but slight change in expression levels is statistically, but not biologically, meaningful.
133 We therefore considered an oligo to have enhancer activity only when (1) the oligo had statistically
134 significant enhancer activity ($p_{FDR}<0.05$) and (2) we observed at least a 50% increase in
135 transcriptional activity compared to the corresponding barcode counts in the plasmid control.
136 Based on these criteria, 16% of SLE risk variants (482 variants, 853 alleles) demonstrated
137 enhancer activity, henceforth referred to as “enhancer variants” (enVars) and “enhancer alleles”
138 (enAlleles), respectively (**Figure 2A, Supplemental Data Set 6**).

139 We next explored the potential effects of enVars on gene expression. We connected each enVar
140 to one or more genes using an approach that takes into account chromatin looping interactions,
141 expression quantitative trait loci (eQTLs), and gene proximity (**Supplemental Data Set 2,**
142 **Supplemental Data Set 7**) (see Methods). This approach identified 1,006 genes in total, which
143 are enriched for expected SLE-related processes such as the interferon pathway, the antigen
144 processing and presentation pathway, and cytokine-related pathways (**Supplemental Figure 3,**
145 **Supplemental Data Set 8**), providing functional support for the enVars we identified.

146 Next, we searched for functional genomic features enriched within enVars relative to non-enVars
147 using the RELI algorithm[10]. In brief, RELI estimates the significance of the intersection between
148 an input set of genomic regions (e.g., enVars) and each member of a collection of functional
149 genomics datasets (e.g., ChIP-seq for a particular histone mark or TF). For this analysis, we
150 identified, curated, and systematically processed the 576 GM12878 ChIP-seq datasets available
151 in the NCBI Gene Expression Omnibus (GEO) database (see Methods). Using RELI, we observed
152 significant enrichment for overlap between enVars and multiple histone modification marks,
153 including H3K4me3 (5.8-fold, $p_{corrected} < 10^{-21}$) and H3K27ac (2.0-fold, $p_{corrected} < 10^{-13}$) (**Figure 2B,**
154 **Supplemental Data Set 9**). As expected, we did not identify enrichment for repressive marks
155 such as H3K9me3 or H3K27me3[28] (**Figure 2B, Supplemental Data Set 9**). Altogether, the
156 genomic features present within enVars confirm that many SLE genetic risk loci likely alter
157 transcriptional regulation in EBV-transformed B cells.

158 We next asked if the genomic binding sites of particular TFs were enriched within our enVars
159 using RELI and the TF ChIP-seq datasets from GM12878. As expected, the enVars are highly
160 enriched for ChIP-seq signal of TFs involved in regulation of the immune response, relative to
161 variants lacking enhancer activity (**Figure 2C**, **Supplemental Data Set 9**). In particular, we found
162 significant enrichment for all members of the NFκB TF family: REL/C-Rel (6.4-fold, $p_{corrected} < 10^{-}$
163 ^{26}), NFKB1/p50 (3.0-fold, $p_{corrected} < 10^{-18}$), RELA/p65 (3.1-fold, $p_{corrected} < 10^{-16}$), RELB (2.7-fold,

164 $p_{corrected} < 10^{-10}$), and NFKB2/p52 (2.2-fold, $p_{corrected} < 10^{-7}$). These results are consistent with our
165 previous findings that altered binding of NFκB TFs is likely an important mechanism conferring
166 SLE risk[10]. We also found significant enrichment for other TFs that have been previously
167 implicated in SLE pathogenesis, such as PAX5[29], MED1[30], IKZF1[31], ELF1[32] and the EBV-encoded
168 EBNA2 transactivator[10] (**Figure 2C, Supplemental Data Set 9**). As a complementary approach,
169 we next assessed enrichment for TF binding site motifs in the enAllele DNA sequences using
170 HOMER[33] and motifs contained in the Cis-BP database[34] (see Methods). This analysis also
171 revealed enrichment of multiple TF families with known roles in SLE, including ETS, NFκB, and
172 IRF[3] (**Figure 2D, Supplemental Data Set 10**). Many of these same TFs also have enriched ChIP-
173 seq peaks at SLE risk loci[10]. Collectively, these results indicate that particular TFs tend to not only
174 concentrate at SLE risk loci[10], but also concentrate at alleles capable of driving gene expression
175 in EBV-transformed B cells.

176

177 **MPRA identifies 51 SLE risk variants with allelic enhancer activity in EBV-transformed B**
178 **cells**

179 We next used our MPRA library to identify SLE genetic risk variants that drive allele-dependent
180 (allelic) enhancer activity. Allelic activity was assessed for each enVar by comparing enhancer
181 activity between each pairs of alleles. We considered a SLE variant allelic if (1) at least one of its
182 alleles is an enAllele; (2) we observed significant genotype-dependent activity using Student's *t*-
183 test[19,35] (see **Supplemental Note 2, Supplemental Figure 6**); and (3) the oligos had more than
184 a 25% change between any pair of alleles. Using these criteria, we identified 51 SLE risk variants
185 (11% of enVars, 1.7% of all SLE risk variants) as allelic enVars in GM12878 (**Figure 3A,**
186 **Supplemental Data Set 11**). For 31 of these 51 allelic enVars, the risk allele decreased enhancer
187 activity relative to the non-risk allele, which is statistically indistinguishable from the 20 variants
188 with increased risk allele activity (p=0.1). Three of the allelic enVars can also alter the amino acid

189 sequence of proteins – rs1059702 (IRAK1), rs1804182 (PLAT), and rs3027878 (HCFC1),
190 consistent with previous studies identifying dual-use codons in the human genome[36]. Collectively,
191 these 51 variants represent causal variant candidates for 27 SLE risk loci (30% of all tested loci)
192 (**Supplemental Data Set 12**). For these 27 risk loci, our approach reduced the number of potential
193 causal variants with allelic activity in GM12878 from an average of 84 variants to an average of
194 two variants per risk locus (**Figure 3B**). For example, at 17q12 (marked by rs8079075), we reduce
195 the candidate causal variant set from 249 to one, with the rs112569955 "G" risk allele showing a
196 36% increase in enhancer activity compared to the "A" non-risk allele.

197

198 **Particular TFs have altered binding at SLE loci with allelic enhancer activity**

199 To identify candidate regulatory proteins that might participate in allelic SLE mechanisms, we next
200 used RELI to identify GM12878 ChIP-seq datasets that significantly overlap allelic enVars
201 (**Supplemental Data Set 13**). Many of the top results are consistent with our previous study[10],
202 including the enriched presence of general enhancer features such as the H3K27ac histone mark
203 (17 of 51 allelic enVars, 13.6-fold enriched, $p_{corrected} < 10^{-38}$), mediator complex subunit MED1 (17
204 of 51 allelic enVars, 13.0-fold enriched, $p_{corrected} < 10^{-34}$), and the histone acetyltransferase p300
205 (16 of 51 allelic enVars, 12.4-fold enriched, $p_{corrected} < 10^{-32}$), along with particular regulatory
206 proteins that participate in "EBV super enhancers"[37] and play key roles in B cells such as ATF7
207 (15 of 51 allelic enVars, 11.3-fold enriched, $p_{corrected} < 10^{-26}$), Ikaros/IKZF1 (19 of 51 allelic enVars,
208 9.7-fold enriched, $p_{corrected} < 10^{-25}$), and the NFκB subunit RELA (13 of 51 allelic enVars, 12.4-fold
209 enriched, $p_{corrected} < 10^{-24}$). Also consistent with our previous study[10], we observe strong enrichment
210 for the EBV-encoded EBNA2 protein (7 of 51 allelic enVars, 17.7-fold enriched, $p_{corrected} < 10^{-19}$).
211 Collectively, these data reveal particular regulatory proteins that might participate in the
212 mechanisms contributing to SLE at multiple risk loci by driving allelic enhancer activity.

213 We next used the MARIO pipeline[10] to search for allelic binding events (i.e. allelic imbalance
214 between sequencing read counts) at SLE variants within 1,058 LCL ChIP-seq datasets (576 from
215 GM12878). By necessity, this approach is limited to the 47 allelic enVars that are heterozygous
216 in at least one of these cell lines. In total, this procedure identified 11 variants with strong allelic
217 imbalance (MARIO ARS value > 0.4) in at least one ChIP-seq dataset (**Supplemental Data Set**
218 **14**), revealing groups of TFs and transcriptional regulators that allelically bind SLE risk variants
219 with genotype-dependent MPRA enhancer activity. For example, the rs3101018 variant, which is
220 associated with SLE[38] and rheumatoid arthritis[39] in Europeans, shows 1.7-fold stronger enhancer
221 activity for the reference/non-risk 'C' allele compared to the non-reference/risk 'T' allele (**Figure**
222 **4A**). These results are consistent with a previously established eQTL obtained from GTEx[40],
223 which demonstrates higher Complement C4A (*C4A*) expression in EBV-transformed B cell lines
224 for the rs3101018 'C' allele than the 'T' allele (**Figure 4B**). Our MARIO allelic ChIP-seq analysis
225 reveals 15 regulatory proteins that prefer the 'C' allele and 2 that prefer the 'T' allele (**Figure 4C,**
226 **Supplemental Figure 4A**). Among these, particularly robust signal is obtained for ATF7, with one
227 experimental replicate in GM12878 displaying 77 vs. 18 reads ('C' vs. 'T') and another showing
228 66 vs. 23 reads ('C' vs. 'T') (**Supplemental Data Set 14**). Moreover, CREB1 and CREM strongly
229 favor the 'C' allele as well (**Supplemental Data Set 14**). In agreement with these data,
230 computational analysis of the DNA sequences surrounding this variant predicts that ATF7,
231 CREB1, and CREM will all bind more strongly to the 'C' than the 'T' allele (**Figure 4D**). Intriguingly,
232 ten additional proteins (FOXK2, PKNOX1, ARID3A, ZBTB40, ZNF217, ARNT, ELF1, IKZF2,
233 MEF2B, and FOXM1) also bind allelically and have known DNA binding motifs[41], but none of them
234 have binding sites altered by the variant. Further, we do not observe allelic chromatin accessibility
235 in an available GM12878 ATAC-seq dataset (9 vs. 7 unique reads). Together, these results reveal
236 a potentially causative SLE regulatory mechanism involving weaker direct binding of
237 ATF7/CREB1/CREM to the 'T' risk allele, altering the recruitment of additional proteins to the
238 locus and lowering the expression of *C4A*.

239 We observe a similar phenomenon for the rs2069235 variant, which is associated with SLE in
240 Asian ancestries[42] and rheumatoid arthritis in Europeans[43]. rs2069235 displays much stronger
241 enhancer activity for the 'A' (non-reference/risk) allele compared to the 'G' (reference/non-risk)
242 allele (**Figure 4E**), consistent with the established synaptogyrin 1 (*SYNGR1)* eQTL in EBV-
243 transformed B cell lines[40] (**Figure 4F**). Inspection of our allelic ChIP-seq data reveals 14 proteins
244 preferentially binding the 'A' allele, with none preferring the 'G' allele (**Figure 4G, Supplemental**
245 **Figure 4B**). Among these 14 proteins, only ELF1 has its binding site directly altered by the variant
246 (**Figure 4H**). Strikingly, 55 of the 78 available H3K27ac datasets are allelic at this variant, with all
247 55 preferring the 'A' allele. Likewise, 24 of 46 H3K4me1 datasets are allelic, with all of them also
248 preferring the 'A' allele. Both of these histone marks are indicative of active chromatin[28]. Only a
249 single histone mark dataset prefers the 'G' allele – the H3K27me3 mark, which is indicative of
250 silenced chromatin[28] (**Figure 4G, Supplemental Figure 4B**). Together, these data are consistent
251 with a potentially causative SLE molecular mechanism involving an allele-dependent enhancer
252 consisting of stronger direct binding of ELF1 to the 'A' risk allele, along with indirectly altered
253 binding of multiple additional TFs to this locus.

254

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

257 As illustrated by the above examples, a particular TF can be involved in allelic mechanisms that
258 are either directly impacted by a given variant (i.e. the variant directly alters the TF's DNA-binding
259 site) or indirectly impacted by the variant (i.e. the variant alters the DNA binding of the TF's
260 physical interaction partner, modulates chromatin accessibility, or affects another mechanism). At
261 a given locus, we designate such TFs as *variant overlapping* and *variant adjacent* TFs,
262 respectively (**Figure 5A**). We next sought to discover such TFs at the 51 allelic enVars. At each
263 allelic enVar locus, we identified *variant overlapping* TFs as those TFs predicted to have strong

264 binding to one allele and weak binding to the other allele. Likewise, we identified *variant adjacent*
265 TFs as those TFs with proximal strong predicted binding sites that do not directly overlap the
266 variant (see Methods). We then searched for particular TFs that tend to act as *variant overlapping*
267 TFs or as *variant adjacent* TFs at the 51 allelic eVars using a proportion test (see Methods) and
268 confirmed that their binding site locations are distributed relative to the variant as expected
269 (**Figure 5B**). Consistent with our results at the *C4A* and *SYNGR1* loci, *variant overlapping* TFs
270 include members of the ETS (e.g., ELF1) and ATF-like (e.g., ATF7) families, along with other TFs
271 whose genetic loci are associated with SLE, including IRF5[44] (**Figure 5C**, **Supplemental Data**
272 **Set 15**). *Variant adjacent* TFs represent a largely distinct class, but also include several TFs with
273 SLE genetic associations, including NFκB[45], the Ikaros (IKZF) family[46], and HMGA family
274 members[47] (**Figure 5D, Supplemental Data Set 15**). Collectively, these analyses reveal two
275 distinct classes of TFs at a given SLE-associated locus that both likely play key roles in SLE
276 mechanisms, along with particular TFs that tend to participate in one class or the other.

277

278 **Allelic transcription regulatory mechanisms shared and unique across cell types**

279 To explore the cell-type specificity of allelic enhancer activity, we transfected our SLE MPRA
280 library into Jurkat cells, a T cell line, as T cells are another important cell type in SLE[48]. We
281 identified 92 SLE risk variants as allelic enVars in Jurkat cells, 25% of which were also found in
282 GM12878 (**Supplemental Figure 5A, B, Supplemental Data Set 11**). We then repeated the
283 experiment in Jurkat cells stimulated with the inflammatory cytokine TNFα, a key cytokine in SLE
284 development[49], to identify stimulation-dependent allelic enVars. This resulted in the identification
285 of 102 allelic enVars, 28 of which were specific to the stimulated Jurkat cells (**Supplemental**
286 **Figure 5C, D, Supplemental Data Set 11**). Altogether, our study identified a total of 145 allelic
287 enVars across 50 independent SLE risk loci (**Supplemental Figure 5E**). These results highlight

288 allelic transcriptional regulatory mechanisms that are both cell-type and inflammatory signaling
289 dependent.

290 In summary, through the application of an allelic MPRA library to the EBV-transformed B cell line
291 GM12878, we identified global transcriptional enhancer activity at 16% of SLE-associated genetic
292 variants (enVars), with particular transcriptional regulatory proteins concentrated at these
293 genomic locations. We further identified 51 SLE risk variants with allelic enhancer activity (allelic
294 enVars) that we now nominate as plausibly causal by acting through genotype-dependent
295 changes to enhancer activity in GM12878. Upon comparison to allelic enhancer activity in the
296 Jurkat T cell line, we identified shared and unique allelic transcriptional regulatory mechanisms at
297 SLE risk loci. Using experimental TF ChIP-seq data and TF binding site motif scanning, we
298 propose a model where the collective action of the genotype-dependent binding of particular
299 *variant overlapping* and *variant adjacent* TFs leads to genotype-dependent transcriptional activity
300 at SLE risk loci.

301

302 **Discussion**

303 Genome-wide association studies identify genetic loci with statistical disease associations.
304 However, each risk locus often contains many plausibly causal variants due to linkage
305 disequilibrium. This study is the first direct genome-wide measurement of enhancer activity at the
306 ~3,000 known SLE genetic risk variants in any context. Unbiased experimental approaches such
307 as MPRA are vital for resolving causal variants and their molecular mechanisms of action.

308 Our results indicate that 16% of the SLE risk variants examined in this study have enhancer
309 activity in the EBV-transformed B cell line GM12878. Furthermore, 51 of these enhancer variants
310 at 27 loci have allelic enhancer activity. These findings are consistent with the theory that a large
311 proportion of the genetic risk of SLE is mediated through transcriptional perturbation of critical B

312 cell genes. Importantly, SLE risk loci exhibit cell-type and inflammatory signaling dependent allelic
313 enhancer activity, with only 25% of allelic enVars shared between GM12878 and Jurkat cells.
314 These results highlight both shared and unique allelic transcriptional regulatory mechanisms for
315 SLE risk.

316 In this study, we used the EBV-transformed B cell line GM12878 as a model for exploring the
317 effects of SLE risk variants. Previous studies have shown that B cells play a critical role in SLE
318 development as immune cells that secrete autoantibodies driving etiology[11]. The relationship
319 between EBV and SLE is widely appreciated. For example, EBV-infected B cells are more
320 prevalent in SLE patients than in healthy people[13,14] and patients with SLE have a higher EBV
321 viral load and infection rate relative to controls[15,16]. Additionally, the EBV transcriptional regulator
322 EBNA2 occupies SLE risk loci in a genotype-dependent manner[10]. *In vivo,* EBV infection can
323 convert primary B cells to activated lymphoblasts[50,51]. EBV will eventually enter latency in resting
324 memory B cells and establish lifelong infection[52]. *In vitro*, EBV infection transforms B cells into
325 immortalized lymphoblastoid cell lines (LCLs)[17]. While we chose the EBV-transformed B cell line
326 GM12878 as the primary disease model for our study, there are limitations to the use of EBV-
327 transformed B cell lines. For example, GM12878 is an immortalized cell line with differences in
328 DNA methylation and gene expression levels from resting and activated B cells[53,54]. A recent study
329 also suggests that EBV infection causes B cells to undergo a germinal center-like differentiation
330 into cells partially resembling plasmablasts and early plasma cells[55], which is only a transient
331 stage *in vivo*. Altogether, these limitations need to be considered when interpreting allelic
332 transcriptional regulation of SLE risk variants in EBV-transformed B cell lines.

333 A critical finding of this study is that SLE risk variants with allelic enhancer activity likely alter the
334 binding of many TFs. Although variants can directly affect the binding of *variant overlapping* TFs
335 via disruption of a DNA-binding site, they can also simultaneously alter the binding of other *variant*
336 *adjacent* TFs, presumably via genomic mechanisms such as altered chromatin accessibility,

337 altered histone marks, indirect TF recruitment through physical interactions, changes in DNA
338 shape, or changes to protein interaction partner DNA binding. This finding corroborates the
339 previously proposed genetic variation-mediation model of motif-dependent and motif-independent
340 TF binding[56-58]. In general, a given TF can be *variant overlapping* at one locus and *variant adjacent*
341 at another, as exemplified by ATF7 (**Figure 4C, G**). Nonetheless, particular TF families tend to
342 act as *variant overlapping* TFs at SLE loci (such as Ets, E-box, and ATF), whereas others tend to
343 act as *variant adjacent* TFs (such as HMGA, Hox, and NFκB). Notably, many of these *variant*
344 *overlapping* and *variant adjacent* TFs are themselves encoded by genetic risk loci associated with
345 SLE (e.g., IRF5[44], NFκB[45], and ETS1[59]), suggesting that there are multiple means through which
346 a particular TF can contribute to disease-based genetic mechanisms. For example, IRF5 targets
347 might be mis-regulated in an SLE patient due to genetic associations in the promoter of *IRF5* that
348 result in altered IRF5 protein levels[44], or by genetic variants located within or adjacent to IRF5
349 binding sites at other genomic loci. It is currently unknown if these TF attributes are shared with
350 other human diseases.

351 This study reveals possible causal genetic mechanisms involving altered binding of particular TFs
352 at two important SLE risk loci. *C4A* is a component of the inflammatory complement pathway that
353 is critical for the appropriate clearance of apoptotic cells[60]. People without *C4A* due to rare,
354 protein-changing mutations are at a greatly increased risk for autoimmune diseases, including
355 type I diabetes and SLE[61]. Further, the risk of developing SLE is 2.62 times higher in subjects
356 with low total *C4A*[60]. Consistent with this observation, the SLE risk allele of rs3101018 at this
357 genetic locus identified in our MPRA is associated with lower *C4A* expression (**Figure 4A**).
358 Moreover, this variant is an eQTL for *C4A* in EBV-transformed B cell lines and whole blood cells,
359 with the risk allele displaying lower *C4A* expression[62] (**Figure 4B, Supplemental Data Set 7**).
360 rs3101018 is located in the Human Leukocyte Antigen (HLA) region of the genome. While many
361 genetic variants at this locus alter amino acid usage in major histocompatibility complex molecules

362 and affect antigen presentation, non-coding genetic variants across the HLA region have also
363 been demonstrated to affect gene expression independently of HLA-type[63-65]. Additionally, the
364 SLE risk locus encoding *SYNGR1* was recently identified in a high-density genotyping study of
365 subjects with Asian ancestry[42] and also increases disease risk for schizophrenia[66], primary biliary
366 cirrhosis[67], and rheumatoid arthritis[68]. *SYNGR1* is an integral membrane protein that is most
367 robustly expressed in neurons of the central nervous system; however, there is measurable
368 transcription and translation of *SYNGR1* in other tissues, including developing B cells[69]. eQTL
369 data from EBV-transformed B cell lines and whole blood cells[62] (**Figure 4F, Supplemental Data**
370 **Set 7**) further support our MPRA-based findings of SLE risk genotype-dependent enhancer
371 activity and gene expression at this locus. Altogether, the results of our MPRA study provide
372 mechanistic insight into these variants through the identification of allelic enVars that facilitate
373 SLE-risk genotype-dependent gene expression.

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

384

385 **Online Methods**

386 **Variant selection and DNA sequence generation**

387 All SLE-associated genetic risk loci reaching genome-wide significance published through March
388 2018 were included in this study[31,38,44,59,70-79]. A total of 91 genetic risk loci were used for linkage
389 disequilibrium (LD) expansion (r^2>0.8) based on 1000 Genomes Data[80] in the ancestry(ies) of the
390 initial genetic association using PLINK[81] (**Supplemental Data Set 1**). All expanded variants were
391 updated to the dbSNP 151 table[82] from the UCSC table browser[83] based on either variant name
392 or genomic location. Unmappable variants were discarded. We also included 20 genetic variants
393 from the Tewhey *et al.*[19] study as positive and negative controls.

394 For single nucleotide polymorphisms, we pulled 170 base pairs (bps) of hg19-flanking DNA
395 sequences for every allele, with the variant located in the center (84 bps upstream and 85 bps
396 downstream of the variant). For the other types of variants (indels), we designed the flanking
397 sequences to ensure that the longest allele has 170 bps. Adapters (15bps) were added to each
398 sequence at either end (5'-ACTGGCCGCTTGACG - [170 bp oligo] - CACTGCGGCTCCTGC-3')
399 to make a 200 bp DNA sequence (**Supplemental Data Set 4**). For all resulting sequences, we
400 created a forward and reverse complement sequence to compensate for possible DNA synthesis
401 errors. A total of 12,478 oligos (3,093 variants, 6,239 alleles) were obtained from Twist Bioscience.

402 **MPRA experiments**

403 ***Library Assembly***

404 For assembly of the MPRA library, we followed the procedure described by Tewhey *et al.* [19] with
405 minor modifications. In brief, we first created the empty vector pGL4.23ΔxbaΔluc from
406 pGL4.23[luc2/minP] using primer Q5_deletion_rev and Q5_deletion_fwd following the
407 manufacturer's instruction of the Q5 Site-Directed Mutagenesis Kit. Then, 20bps barcodes were
408 added to the synthesized oligos through 24X PCR with 50 μL system, each containing 1.86ng
409 oligo, 25 μL NEBNext® Ultra™ II Q5® Master Mix, 1μM MPRA_v3_F and MPRA_v3_201_R.

410 PCR was performed under the following conditions: 98°C for 2 mins, 12 cycles of (98°C for 10
411 sec, 60°C for 15 sec, 72°C for 45 sec), 72°C for 5 mins. Amplified product was purified and cloned
412 into SfiI digested pGL4.23ΔxbaΔluc by Gibson assembly at 50°C for 1 hr. The assembled
413 backbone library was purified and then transformed into *Escherichia coli* (*E.coli)* through
414 electroporation (2kV, 200 ohm, 25 μF). Electroporated *E.coli* was expanded in 200 mL of LB Broth
415 buffer supplemented with 100 μg/mL of carbenicillin at 37°C for 12 to 16 hrs. Plasmid was then
416 extracted using the QIAGEN Plasmid Maxi Kit.

417 We next created the pGL4.23[eGFP/miniP] plasmid. An eGFP fragment was amplified from MS2-
418 P65-HSF1_GFP (Addgene #61423) through PCR with a 50μl system containing 1 ng plasmid, 25
419 μL NEBNext® Ultra™ II Q5® Master Mix, 0.5 μM GFP_seq_MS2-P65-HSF1_GFP_FWD and
420 GFP_seq_MS2-P65-HSF1_GFP_REV. PCR was performed under the following conditions: 98°C
421 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
422 amplified fragment was purified and then inserted into XbaI and NcoI digested pGL4.23[luc2/minP]
423 through Gibson assembly at 50°C for 1 hr. The assembled plasmid was purified and then
424 transformed into *E.coli* through chemical transformation. Transformed *E.coli* was expanded in 100
425 mL of LB Broth buffer supplemented with 100 μg/mL of carbenicillin at 37°C for 12 to 16 hrs.
426 Plasmid was then extracted using the QIAGEN Plasmid Maxi Kit.

427 A miniP + eGFP fragment was amplified from pGL4.23[eGFP/miniP] through 8X PCR with 50 μL
428 system, each containing 1 ng plasmid, 25 μL NEBNext® Ultra™ II Q5® Master Mix, 0.5μM 200-
429 MPRA_v3_GFP_Fusion_v2_F and 201-MPRA_v3_GFP_Fusion_v2_R. PCR was performed
430 under the following conditions: 98°C for 2 mins, 20 cycles of (98°C for 10 sec, 60°C for 15 sec,
431 72°C for 45 sec), 72°C for 5 mins. The amplified product was purified and then inserted into AsiSI
432 digested backbone library through Gibson assembly at 50°C for 1.5 hrs to create the transfection
433 library. The resulting library was re-digested by RecBCD and AsiSI, purified and then transformed

434 into *E.coli* through electroporation (2kV, 200 ohm, 25 μF). Transformed *E.coli* was cultured in 5L
435 of LB Broth buffer supplemented with 100 μg/mL of carbenicillin at 37 °C for 12 to 16 hrs. The
436 plasmid was then extracted using the QIAGEN Endo-free Plasmid Giga Kit.

437 ***Oligo and barcode association***

438 The oligo and barcode regions were amplified from the backbone library through 4X PCR with a
439 100 μL system containing 200 ng plasmid, 50 μL NEBNext® Ultra™ II Q5® Master Mix, 0.5 μM
440 TruSeq_Universal_Adapter_P5 and MPRA_v3_TruSeq_Amp2Sa_F_P7. PCR was performed
441 under the following conditions: 95°C for 20 sec, 6 cycles of (95°C for 20 sec, 62°C for 15 sec,
442 72°C for 30 sec), 72°C for 2 mins. The product was then purified, and indices were added through
443 a 100μl system containing all purified product, 50μl NEBNext® Ultra™ II Q5® Master Mix, 0.5 μM
444 TruSeq_Universal_Adapter_P5 and index primer. PCR was performed as above, except for only
445 5 cycles. Samples were purified, molar pooled, and sequenced using 2x125bp on Illumina
446 NextSeq 500.

447 ***Transfection***

448 The GM12878 cell line was grown in RPMI medium supplemented with 10% FBS, 100 units/mL
449 of penicillin, and 100 μg/mL of streptomycin. Cells were seeded at a density of 5 x 10^5 cells/mL
450 the day before transfection. For triplicate transfections, we collected a total of 5 x 10^7 cells per
451 replicate. Cells were then suspended with 50 μg transfection library plasmid in 400 μL Buffer R.
452 Electroporation was performed with the Neon transfection system in 100μl tips with 3 pulses of
453 1200V, 20 ms each. After transfection, cells were recovered in 50 mL pre-warmed RPMI medium
454 supplemented only with 10% FBS for 24 hrs. Cells were then collected for preparation of the
455 sequencing library for barcode counting.

456 The Jurkat cell line was grown in RPMI medium supplemented with 10% FBS, 100 units/mL of
457 penicillin, and 100 μg/mL of streptomycin. Cells were seeded at a density of 5 x 10^5 cells/mL the

458 day before transfection. For each experimental group, we collected a total of 5 x 10^7 cells per
459 replicate for 5 replicates. Cells were then resuspended with 50 μg transfection library plasmid in
460 400 μL Buffer R. Electroporation was performed with the Neon transfection system in 100μl tips
461 with 3 pulses of 1350V, 10 ms each. After transfection, cells were recovered in 50 mL pre-warmed
462 RPMI medium supplemented only with 10% FBS for 24 hrs. After recovery, cells were
463 supplemented with or without 100ng/ml TNFα for 24 hrs. Cells were then collected for preparation
464 of the sequencing library for barcode counting.

465 ***Sequencing library for barcode counting***

466 For sample from GM12878 cell, total RNA of transfected cells was extracted by the RNeasy Midi
467 Kit following the manufacturer's instruction. Extracted RNA was subjected to DNase treatment in
468 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
469 μL 10% SDS and 37.5 μL 0.5M EDTA were added to stop DNase with 5 mins of incubation at
470 75°C. The whole volume was used for eGFP probe hybridization in an 1800 μL system, with 450
471 μl 20X SSC Buffer, 900 μL Formamide and 1 μL of each 100 μM Biotin-labeled GFP probe One
472 to Three. The probe hybridization was performed through incubation at 65°C for 2.5 hrs. 200 μL
473 Dynabeads™ MyOne™ Streptavidin C1 was prepared according to the manufacturer's instruction.
474 The beads were suspended in 250 μL 20X SSC Buffer and incubated with the above probe
475 hybridization reaction at room temperature for 15 mins. Beads were then collected on a magnet
476 and washed with 1X SSC Buffer once, and 0.1X SSC Buffer twice. eGFP mRNA was eluted first
477 through adding 12.5 μL ddH2O, heating at 70°C for 2 mins and collecting on a magnet, then
478 adding another 12.5 μL ddH2O, heating at 80°C for 2 mins and collecting on a magnet. All
479 collected elution was performed with another DNase treatment in a 30 μL system containing 0.5
480 μL Turbo DNase and 3 μL Turbo DNase Buffer at 37°C for 1 hr. 0.5 μL 10% SDS was added to
481 halt DNase reaction. Eluted mRNA was purified through RNA Clean SPRI Beads. mRNA was

482 reverse transcribed to cDNA using SuperScript™ IV First-Strand Synthesis System with gene
483 specific primer MPRA_v3_Amp2Sc_R, following the manufacturer's instruction. cDNA and
484 plasmid control were then used for building sequencing libraries following the Tag-seq Library
485 Construction section in the paper of Tewhey *et al.*[19]. A total of two PCRs were needed for building
486 the sequencing library. The first PCR was performed with TruSeq_Universal_Adapter_P5 and
487 MPRA_V3_Illumina_GFP_F. The second PCR was performed with
488 TruSeq_Universal_Adapter_P5 and index primer. Samples were purified, molar pooled, and
489 sequenced using 1x75bp on Illumina NextSeq 500.

490 For samples from Jurkat cells, total DNA and RNA of transfected cells were extracted by the
491 Qiagen ALLPrep DNA/RNA Mini Kit following the manufacturer's instruction[23]. Extracted RNA was
492 processed the same as above to obtain cDNA. cDNA, extracted DNA, and plasmid control were
493 then used for building sequencing libraries following the Tag-seq Library Construction section in
494 the paper of Tewhey *et al.*[19]. A total of two PCRs were needed for building the sequencing library.
495 The first PCR was performed with TruSeq_Universal_Adapter_P5 and
496 MPRA_V3_Illumina_GFP_F. The second PCR was performed with
497 TruSeq_Universal_Adapter_P5 and index primer. Samples were purified, molar pooled, and
498 sequenced using 1x100bp on Illumina NovaSeq 6000.

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

500 **MPRA data analysis**

501 ***Oligo and barcode association***

502 Paired-end, 125 bp reads were first quality filtered using Trimmomatic-0.38[84] (flags: PE -phred33,
503 LEADING:25, TRAILING:25, MINLEN:80). Read 1 was then separated into the 20bp barcode
504 region and the oligo-matching region. The trimmed oligo-matching regions of Read 1 and Read 2
505 were mapped back to the synthesized oligo sequences using Bowtie2[85] (flags: -X 250, --very-

506 sensitive, -p 16). Barcodes were then associated with the oligo sequences using the read ID. Only
507 uniquely mapped barcodes were used for downstream analysis.

508 ***Barcode counting***

509 Single end 75/100 bp reads were first quality filtered using Trimmomatic-0.38[84] (flags: PE -
510 phred33, LEADING:3, TRAILING:3, MINLEN:70). Each read was then separated into the 20bp
511 barcode region and the constant region. The trimmed constant regions of the reads were mapped
512 back to the constant region within the eGFP 3' UTR using Bowtie2[85] (flags: --very-sensitive, -p
513 16). Only reads with Levenshtein distance of 4 or less within the constant region and perfect
514 matches to the two bases directly adjacent to the barcode were kept. Barcodes were then
515 associated with the retained reads using the read ID. Only barcodes that met our quality threshold
516 requirements described above in the methods section "*Oligo and barcode association"* were used
517 for downstream analysis.

518 ***Enhancer variant (EnVar) identification***

519 We followed the procedures described in the "Identification of Regulatory Oligos" section of
520 Tewhey *et al.*[19] with minor modifications. In brief, oligos (alleles) with 30 or more unique barcodes
521 from the plasmid control were included for analysis. All barcodes were summarized at the oligo
522 level. Barcode count totals for each oligo, including all SLE variants and the 20 control variants,
523 were passed into DESeq2[86] (R version 3.5.3[87]) to estimate the fold change and significance
524 between plasmid controls (see **Supplemental Note 1, Supplemental Figure 5F, G**) and the
525 experiment replicates. A Benjamini-Hochberg FDR adjusted p-value of less than 0.05 was
526 required for significance. Only significant alleles with greater than or equal to a 1.5x fold change
527 were identified as enhancer alleles (enAlleles). A variant was identified as an enhancer variant
528 (enVar) if any allele of this variant was an enAllele. Results for the 20 control variants were
529 compared to data from Tewhey *et al.*[19] to estimate accuracy, sensitivity, and specificity.

530 ***Allelic enVar identification***

531 Only enVars were considered for allelic analysis. The barcode counts from every allele of each
532 enVar were used for calculating p-values by comparing the log2 ratios of the non-reference allele
533 vs the reference allele, normalized by plasmid controls, using Student's *t*-test[19,35]. p-values were
534 adjusted with the Benjamini-Hochberg FDR-based procedure. A corrected p-value of less than
535 0.05 was required for significance. Only significant alleles with 25%-fold changes or greater were
536 identified as allelic enVars (see **Supplemental Note 2, Supplemental Figure 6**). We have
537 created an R package (mpraprofiler) for performing this analysis, which is available on the
538 Weirauch lab GitHub page (https://github.com/WeirauchLab/).

539 **Downstream MPRA analysis**

540 ***Gene annotation***

541 We annotated each SLE genetic variant with its nearest gene using the NCBI RefSeq table[88]
542 downloaded from the UCSC table browser[83]. enVars were annotated using a combination of DNA
543 looping interactions (GM12878 Capture Hi-C data[89,90]) and eQTL data obtained from the eQTL
544 Catalogue, a resource that contains quality controlled, uniformly re-computed eQTLs from 19
545 eQTL publications[62,91-110], EBV-transformed B cell lines (GTEx Analysis V7 (dbGaP Accession
546 phs000424.v7.p2)[40] and other individual studies[111-114]. For all variants, the target genes were
547 annotated (**Supplemental Data Set 2**) using the union of promoter interacting genes and eQTL
548 genes from B cells with and without EBV transformation, when available. Otherwise, target genes
549 were annotated as the nearest gene. Allelic enVar gene targets were classified into four tiers: a
550 Tier 1) variant is both an eQTL and also loops to the promoter of the same gene; a Tier 2) variant
551 has an eQTL for at least one gene; a Tier 3) variant only loops to the promoter of at least one
552 gene; a Tier 4) variant is neither an eQTL nor loops to the promoter of any gene (**Supplemental**
553 **Data Set 12**).

554 ***TF binding site motif enrichment analysis***

555 To identify specific TFs whose binding might contribute to the enhancer activity observed in our
556 MPRA experiments, we performed HOMER[33] TF binding site motif enrichment analysis.
557 Specifically, we used HOMER to calculate the enrichment of each motif in the sequence of
558 enAlleles compared to the sequences of non-enAlleles. HOMER was modified to use the large
559 library of human position weight matrix (PWM) binding site models contained in build 2.0 of the
560 CisBP database[34] and a log base 2 likelihood scoring system.

561 ***GO enrichment analysis***

562 Enrichr[115,116] was used for GO enrichment analysis. In short, the target genes of enVars were
563 passed to Enrichr for analysis. Results from the GO biological process (2018) category were used
564 (**Supplemental Data Set 8, Supplemental Figure 3)**.

565 ***Identification and processing of publicly available LCL ChIP-seq data***

566 1,058 ChIP-seq datasets were obtained from the Gene Expression Omnibus (GEO)[117] using
567 custom scripts that searched for ChIP-seq experiments performed in EBV-transformed
568 lymphoblastoid cell lines (LCLs). The annotations for every dataset (assay type, cell line, assayed
569 molecule) were manually checked by two authors (MTW and LCK) to ensure accuracy. The
570 Sequence Read Archive (SRA) files obtained from GEO were analyzed using an automated
571 pipeline. Briefly, the pipeline first runs QC on the FastQ files containing the sequencing reads
572 using FastQC (v0.11.2)[118]. If FastQC detects adapter sequences, the pipeline runs the FastQ files
573 through Trim Galore (v0.4.2)[119], a wrapper script that runs cutadapt (v1.9.1)[120] to remove the
574 detected adapter sequence from the reads. The quality controlled reads are then aligned to the
575 reference human genome (hg19/GRCh37) using bowtie2 (v2.3.4.1)[85]. The aligned reads (in .BAM
576 format) are then sorted using samtools (v1.8.0)[121] and duplicate reads are removed using picard
577 (v1.89)[122]. Finally, peaks are called using MACS2 (v2.1.2) (flags: callpeak -g hs -q 0.01 -f BAM)[123].

578 ENCODE blacklist regions[124] were removed from the peak sets using the hg19-blacklist.bed.gz
579 file available at https://github.com/Boyle-Lab/Blacklist/tree/master/lists/Blacklist_v1. ChIP-seq
580 datasets GSM1666207, GSM2748907 and GSM1599157 were removed due to the low number
581 of cells used in the experiments.

582 ***Functional genomics dataset enrichment analysis with RELI***

583 We used the RELI[10] algorithm to identify genomic features (TF binding events, histone marks,
584 etc.) that coincide with enVars. As input, RELI takes the genomic coordinates of enVars. RELI
585 then systematically intersects these coordinates with one of the GM12878 ChIP-seq datasets,
586 and the number of input regions overlapping the peaks of this dataset (by at least one base) is
587 counted. Next, a p-value describing the significance of this overlap is estimated using a
588 simulation-based procedure. To this end, a 'negative set' is created for comparison to the input
589 set, which in this study contains the set of non-enVars (i.e., variants with no allele having an
590 adjusted p-value of less than 0.05 and more than 10%- fold change in the DESeq2 result). A
591 distribution of expected overlap values is then created from 2,000 iterations of randomly sampling
592 from the negative set, each time choosing a set of negative examples that match the input set in
593 terms of the total number of genomic loci. The distribution of the expected overlap values from
594 the randomized data resembles a normal distribution and can thus be used to generate a Z-score
595 and corresponding p-value estimating the significance of the observed number of input regions
596 that overlap each ChIP-seq dataset.

597 We performed similar RELI analysis for allelic enVars. As input, we used the allelic enVar sites.
598 For the 'negative set', we used the set of common SNPs taken from the dbSNP142 database
599 downloaded from the UCSC table browser[83].

600 ***Identification of allelic ChIP-seq reads using MARIO***

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

614 ***Identification of* variant overlapping *and* variant adjacent *TFs***

615 *Variant overlapping* TFs were identified using an algorithm that compares predicted TF binding
616 motif scores between the different alleles of each allelic enVar. First, we padded each allele of a
617 given allelic enVar with 25 bps of upstream and downstream DNA sequence (a sufficient length
618 to account for any known human TF binding sites[125]). The algorithm consists of two major
619 components: (1) individually scoring the two alleles of a given variant with a given TF model; and
620 (2) quantifying the difference in the binding intensity between these two alleles. DNA sequences
621 are scored using the large collection of human TF position frequency matrix (PFM) models
622 contained in the Cis-BP database[34] and the log-likelihood PFM scoring system[126]. Since log-
623 likelihood score distributions vary substantially (depending on the information content of a given
624 motif), we employ a simple scaled scoring system that maps a given log-likelihood score to the
625 percentage of the maximum achievable log-likelihood score of the given motif – we refer to this

626 value as the “relative PFM score”. We identify binding site altering events (i.e. “creating” or
627 “breaking” a predicted binding site for a given TF motif) as cases where one allele has a relative
628 PFM score of 70% or higher, and the other allele has a score of less than 40%. For a given variant,
629 any TF with allelic ChIP-seq sequencing reads (see above) and a binding site altering event for
630 any of its motifs was deemed a *variant overlapping* TF. Any TF with allelic ChIP-seq sequencing
631 reads and a lack of a binding site altering event for any of its motifs was deemed a *variant adjacent*
632 TF.

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

639 We used a similar procedure to identify particular TFs that tend to be *variant adjacent* TFs at SLE
640 allelic enVars. We performed HOMER motif enrichment analysis using the full 170bp allelic enVar
641 DNA sequences as input. The standard HOMER null model (scrambled input sequence,
642 maintaining dinucleotide frequencies) was used as background. The fractions of motif “hits”
643 obtained in the foreground vs. background set were then compared, and significance was again
644 calculated using a proportions test. Results are provided in **Supplemental Data Set 15**.

645

646 **Data availability**

MPRA data are available in the Gene Expression Omnibus (GEO) database under accession number GSE143792. Full datasets and processed results are provided in the Supplementary Material.

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 (F32 AI129249, K99-HG009920, P30 AR070549, P30 DK078392, R01 AI024717, R01 AI148276, R01 AR073228, R01 DK107502, R01 GM055479, R01 HG010730, R01 NS099068, U01 AI130830, U01 HG008666*),* 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.

677 **References**

678 1. Carter, E.E., Barr, S.G. & Clarke, A.E. The global burden of SLE: prevalence, health
679 disparities and socioeconomic impact. *Nat Rev Rheumatol* **12**, 605-20 (2016).
680 2. Tsokos, G.C. Systemic Lupus Erythematosus. *New England Journal of Medicine* **365**,
681 2110-2121 (2011).
682 3. Deng, Y. & Tsao, B.P. Genetic susceptibility to systemic lupus erythematosus in the
683 genomic era. *Nat Rev Rheumatol* **6**, 683-92 (2010).
684 4. Visscher, P.M. *et al.* 10 Years of GWAS Discovery: Biology, Function, and Translation.
685 *Am J Hum Genet* **101**, 5-22 (2017).
686 5. Fike, A.J., Elcheva, I. & Rahman, Z.S.M. The Post-GWAS Era: How to Validate the
687 Contribution of Gene Variants in Lupus. *Curr Rheumatol Rep* **21**, 3 (2019).
688 6. Corradin, O. *et al.* Modeling disease risk through analysis of physical interactions
689 between genetic variants within chromatin regulatory circuitry. *Nat Genet* **48**, 1313-1320
690 (2016).
691 7. Farh, K.K. *et al.* Genetic and epigenetic fine mapping of causal autoimmune disease
692 variants. *Nature* **518**, 337-43 (2015).
693 8. Teruel, M. & Alarcon-Riquelme, M.E. The genetic basis of systemic lupus
694 erythematosus: What are the risk factors and what have we learned. *J Autoimmun* **74**,
695 161-175 (2016).
696 9. Hu, X. *et al.* Integrating autoimmune risk loci with gene-expression data identifies
697 specific pathogenic immune cell subsets. *Am J Hum Genet* **89**, 496-506 (2011).
698 10. Harley, J.B. *et al.* Transcription factors operate across disease loci, with EBNA2
699 implicated in autoimmunity. *Nat Genet* **50**, 699-707 (2018).
700 11. Karrar, S. & Cunninghame Graham, D.S. Abnormal B Cell Development in Systemic
701 Lupus Erythematosus: What the Genetics Tell Us. *Arthritis Rheumatol* **70**, 496-507
702 (2018).
703 12. Dorner, T., Giesecke, C. & Lipsky, P.E. Mechanisms of B cell autoimmunity in SLE.
704 *Arthritis Res Ther* **13**, 243 (2011).
705 13. Draborg, A.H., Duus, K. & Houen, G. Epstein-Barr virus and systemic lupus
706 erythematosus. *Clin Dev Immunol* **2012**, 370516 (2012).
707 14. McClain, M.T. *et al.* Early events in lupus humoral autoimmunity suggest initiation
708 through molecular mimicry. *Nat Med* **11**, 85-9 (2005).
709 15. James, J.A. *et al.* An increased prevalence of Epstein-Barr virus infection in young
710 patients suggests a possible etiology for systemic lupus erythematosus. *J Clin Invest*
711 **100**, 3019-26 (1997).
712 16. Hanlon, P., Avenell, A., Aucott, L. & Vickers, M.A. Systematic review and meta-analysis
713 of the sero-epidemiological association between Epstein-Barr virus and systemic lupus
714 erythematosus. *Arthritis Res Ther* **16**, R3 (2014).
715 17. Pope, J.H. Establishment of Cell Lines from Peripheral Leucocytes in Infectious
716 Mononucleosis. *Nature* **216**, 810-811 (1967).
717 18. Patwardhan, R.P. *et al.* Massively parallel functional dissection of mammalian enhancers
718 in vivo. *Nat Biotechnol* **30**, 265-70 (2012).
719 19. Tewhey, R. *et al.* Direct Identification of Hundreds of Expression-Modulating Variants
720 using a Multiplexed Reporter Assay. *Cell* **165**, 1519-1529 (2016).
721 20. Ulirsch, J.C. *et al.* Systematic Functional Dissection of Common Genetic Variation
722 Affecting Red Blood Cell Traits. *Cell* **165**, 1530-1545 (2016).
723 21. Liu, Y. *et al.* Functional assessment of human enhancer activities using whole-genome
724 STARR-sequencing. *Genome Biol* **18**, 219 (2017).

725 22. Vockley, C.M. *et al.* Massively parallel quantification of the regulatory effects of
726 noncoding genetic variation in a human cohort. *Genome Res* **25**, 1206-14 (2015).
727 23. Klein, J.C. *et al.* Functional testing of thousands of osteoarthritis-associated variants for
728 regulatory activity. *Nat Commun* **10**, 2434 (2019).
729 24. Choi, J. *et al.* Massively parallel reporter assays of melanoma risk variants identify MX2
730 as a gene promoting melanoma. *Nat Commun* **11**, 2718 (2020).
731 25. Lin, X. *et al.* FAM13A Represses AMPK Activity and Regulates Hepatic Glucose and
732 Lipid Metabolism. *iScience* **23**, 100928 (2020).
733 26. Madan, N. *et al.* Functionalization of CD36 cardiovascular disease and expression
734 associated variants by interdisciplinary high throughput analysis. *PLoS Genet* **15**,
735 e1008287 (2019).
736 27. Zhang, S. *et al.* Allele-specific open chromatin in human iPSC neurons elucidates
737 functional disease variants. *Science* **369**, 561-565 (2020).
738 28. Zhou, V.W., Goren, A. & Bernstein, B.E. Charting histone modifications and the
739 functional organization of mammalian genomes. *Nat Rev Genet* **12**, 7-18 (2011).
740 29. Ruer-Laventie, J. *et al.* Overexpression of Fkbp11, a feature of lupus B cells, leads to B
741 cell tolerance breakdown and initiates plasma cell differentiation. *Immun Inflamm Dis* **3**,
742 265-79 (2015).
743 30. Armstrong, D.L. *et al.* GWAS identifies novel SLE susceptibility genes and explains the
744 association of the HLA region. *Genes Immun* **15**, 347-54 (2014).
745 31. Han, J.W. *et al.* Genome-wide association study in a Chinese Han population identifies
746 nine new susceptibility loci for systemic lupus erythematosus. *Nat Genet* **41**, 1234-7
747 (2009).
748 32. Yang, J. *et al.* ELF1 is associated with systemic lupus erythematosus in Asian
749 populations. *Hum Mol Genet* **20**, 601-7 (2011).
750 33. Heinz, S. *et al.* Simple combinations of lineage-determining transcription factors prime
751 cis-regulatory elements required for macrophage and B cell identities. *Mol Cell* **38**, 576-
752 89 (2010).
753 34. Weirauch, M.T. *et al.* Determination and inference of eukaryotic transcription factor
754 sequence specificity. *Cell* **158**, 1431-1443 (2014).
755 35. Ray, J.P. *et al.* Prioritizing disease and trait causal variants at the TNFAIP3 locus using
756 functional and genomic features. *Nat Commun* **11**, 1237 (2020).
757 36. Stergachis, A.B. *et al.* Exonic transcription factor binding directs codon choice and
758 affects protein evolution. *Science* **342**, 1367-72 (2013).
759 37. Zhao, B. *et al.* Epstein-Barr virus exploits intrinsic B-lymphocyte transcription programs
760 to achieve immortal cell growth. *Proc Natl Acad Sci U S A* **108**, 14902-7 (2011).
761 38. International Consortium for Systemic Lupus Erythematosus, G. *et al.* Genome-wide
762 association scan in women with systemic lupus erythematosus identifies susceptibility
763 variants in ITGAM, PXK, KIAA1542 and other loci. *Nat Genet* **40**, 204-10 (2008).
764 39. Black, M.H. & Watanabe, R.M. A principal-components-based clustering method to
765 identify multiple variants associated with rheumatoid arthritis and arthritis-related
766 autoantibodies. *BMC Proc* **3 Suppl 7**, S129 (2009).
767 40. Consortium, G.T. The Genotype-Tissue Expression (GTEx) project. *Nat Genet* **45**, 580-5
768 (2013).
769 41. Lambert, S.A. *et al.* Similarity regression predicts evolution of transcription factor
770 sequence specificity. *Nat Genet* **51**, 981-989 (2019).
771 42. Sun, C. *et al.* High-density genotyping of immune-related loci identifies new SLE risk
772 variants in individuals with Asian ancestry. *Nat Genet* **48**, 323-30 (2016).
773 43. Okada, Y. *et al.* Genetics of rheumatoid arthritis contributes to biology and drug
774 discovery. *Nature* **506**, 376-81 (2014).

775 44. Kottyan, L.C. *et al.* The IRF5-TNPO3 association with systemic lupus erythematosus has
776 two components that other autoimmune disorders variably share. *Hum Mol Genet* **24**,
777 582-96 (2015).
778 45. Li, Y. *et al.* Genetic Variants of IkappaB Kinase beta (IKBKB) and Polymerase beta
779 (POLB) Were Not Associated with Systemic Lupus Erythematosus Risk in a Chinese
780 Han Population. *PLoS One* **10**, e0132556 (2015).
781 46. Zhang, Y.M. *et al.* Association of the IKZF1 5' UTR variant rs1456896 with lupus
782 nephritis in a northern Han Chinese population. *Scand J Rheumatol* **46**, 210-214 (2017).
783 47. Patel, Z.H. *et al.* A plausibly causal functional lupus-associated risk variant in the
784 STAT1-STAT4 locus. *Hum Mol Genet* **27**, 2392-2404 (2018).
785 48. Tsokos, G.C. Systemic lupus erythematosus. *N Engl J Med* **365**, 2110-21 (2011).
786 49. Zhu, L.J., Yang, X. & Yu, X.Q. Anti-TNF-alpha therapies in systemic lupus
787 erythematosus. *J Biomed Biotechnol* **2010**, 465898 (2010).
788 50. Zhou, H. *et al.* Epstein-Barr Virus Oncoprotein Super-enhancers Control B Cell Growth.
789 *Cell Host & Microbe* **17**, 205-216 (2015).
790 51. Kurth, J. *et al.* EBV-Infected B Cells in Infectious Mononucleosis: Viral Strategies for
791 Spreading in the B Cell Compartment and Establishing Latency. *Immunity* **13**, 485-495
792 (2000).
793 52. Babcock, G.J., Hochberg, D. & Thorley-Lawson, D.A. The Expression Pattern of
794 Epstein-Barr Virus Latent Genes In Vivo Is Dependent upon the Differentiation Stage of
795 the Infected B Cell. *Immunity* **13**, 497-506 (2000).
796 53. Hernando, H. *et al.* The B cell transcription program mediates hypomethylation and
797 overexpression of key genes in Epstein-Barr virus-associated proliferative conversion.
798 *Genome Biology* **14**, R3 (2013).
799 54. Hansen, K.D. *et al.* Large-scale hypomethylated blocks associated with Epstein-Barr
800 virus-induced B-cell immortalization. *Genome Res* **24**, 177-84 (2014).
801 55. Mrozek-Gorska, P. *et al.* Epstein–Barr virus reprograms human B lymphocytes
802 immediately in the prelatent phase of infection. *Proceedings of the National Academy of*
803 *Sciences* **116**, 16046 (2019).
804 56. Deplancke, B., Alpern, D. & Gardeux, V. The Genetics of Transcription Factor DNA
805 Binding Variation. *Cell* **166**, 538-554 (2016).
806 57. Kilpinen, H. *et al.* Coordinated effects of sequence variation on DNA binding, chromatin
807 structure, and transcription. *Science* **342**, 744-7 (2013).
808 58. Reddy, T.E. *et al.* Effects of sequence variation on differential allelic transcription factor
809 occupancy and gene expression. *Genome Res* **22**, 860-9 (2012).
810 59. Lu, X. *et al.* Lupus Risk Variant Increases pSTAT1 Binding and Decreases ETS1
811 Expression. *Am J Hum Genet* **96**, 731-9 (2015).
812 60. Tsang, A.S.M.W.P. *et al.* Comprehensive approach to study complement C4 in systemic
813 lupus erythematosus: Gene polymorphisms, protein levels and functional activity. *Mol*
814 *Immunol* **92**, 125-131 (2017).
815 61. Juptner, M. *et al.* Low copy numbers of complement C4 and homozygous deficiency of
816 C4A may predispose to severe disease and earlier disease onset in patients with
817 systemic lupus erythematosus. *Lupus* **27**, 600-609 (2018).
818 62. Momozawa, Y. *et al.* IBD risk loci are enriched in multigenic regulatory modules
819 encompassing putative causative genes. *Nat Commun* **9**, 2427 (2018).
820 63. D'Antonio, M. *et al.* Systematic genetic analysis of the MHC region reveals mechanistic
821 underpinnings of HLA type associations with disease. *Elife* **8**(2019).
822 64. Lam, T.H., Shen, M., Tay, M.Z. & Ren, E.C. Unique Allelic eQTL Clusters in Human
823 MHC Haplotypes. *G3 (Bethesda)* **7**, 2595-2604 (2017).

824 65. Aguiar, V.R.C., Cesar, J., Delaneau, O., Dermitzakis, E.T. & Meyer, D. Expression
825 estimation and eQTL mapping for HLA genes with a personalized pipeline. *PLoS Genet*
826 **15**, e1008091 (2019).
827 66. Iatropoulos, P. *et al.* Association study and mutational screening of SYNGR1 as a
828 candidate susceptibility gene for schizophrenia. *Psychiatr Genet* **19**, 237-43 (2009).
829 67. Kim, K. *et al.* High-density genotyping of immune loci in Koreans and Europeans
830 identifies eight new rheumatoid arthritis risk loci. *Ann Rheum Dis* **74**, e13 (2015).
831 68. Liu, J.Z. *et al.* Dense fine-mapping study identifies new susceptibility loci for primary
832 biliary cirrhosis. *Nat Genet* **44**, 1137-41 (2012).
833 69. Wu, C. *et al.* BioGPS: an extensible and customizable portal for querying and organizing
834 gene annotation resources. *Genome Biol* **10**, R130 (2009).
835 70. Liu, L. *et al.* Genome-wide association study identifies three novel susceptibility loci for
836 systemic lupus erythematosus in Han Chinese. *Br J Dermatol* **179**, 506-508 (2018).
837 71. Morris, D.L. *et al.* Genome-wide association meta-analysis in Chinese and European
838 individuals identifies ten new loci associated with systemic lupus erythematosus. *Nat*
839 *Genet* **48**, 940-946 (2016).
840 72. Alarcon-Riquelme, M.E. *et al.* Genome-Wide Association Study in an Amerindian
841 Ancestry Population Reveals Novel Systemic Lupus Erythematosus Risk Loci and the
842 Role of European Admixture. *Arthritis Rheumatol* **68**, 932-43 (2016).
843 73. Zhang, Y. *et al.* Genome-wide search followed by replication reveals genetic interaction
844 of CD80 and ALOX5AP associated with systemic lupus erythematosus in Asian
845 populations. *Ann Rheum Dis* **75**, 891-8 (2016).
846 74. Kunz, M. *et al.* Genome-wide association study identifies new susceptibility loci for
847 cutaneous lupus erythematosus. *Exp Dermatol* **24**, 510-5 (2015).
848 75. Lei, S.F. & Deng, F.Y. Identification of susceptibility genes for systemic lupus
849 erythematosus with a genome-wide gene-based association study. *Scand J Rheumatol*
850 **43**, 426-8 (2014).
851 76. Okada, Y. *et al.* A genome-wide association study identified AFF1 as a susceptibility
852 locus for systemic lupus eyrthematosus in Japanese. *PLoS Genet* **8**, e1002455 (2012).
853 77. Yang, W. *et al.* Genome-wide association study in Asian populations identifies variants
854 in ETS1 and WDFY4 associated with systemic lupus erythematosus. *PLoS Genet* **6**,
855 e1000841 (2010).
856 78. Cervino, A.C., Tsinoremas, N.F. & Hoffman, R.W. A genome-wide study of lupus:
857 preliminary analysis and data release. *Ann N Y Acad Sci* **1110**, 131-9 (2007).
858 79. Langefeld, C.D. *et al.* Transancestral mapping and genetic load in systemic lupus
859 erythematosus. *Nat Commun* **8**, 16021 (2017).
860 80. The Genomes Project, C. *et al.* A global reference for human genetic variation. *Nature*
861 **526**, 68 (2015).
862 81. Chang, C.C. *et al.* Second-generation PLINK: rising to the challenge of larger and richer
863 datasets. *Gigascience* **4**, 7 (2015).
864 82. Sherry, S.T. *et al.* dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res* **29**,
865 308-11 (2001).
866 83. Karolchik, D. *et al.* The UCSC Table Browser data retrieval tool. *Nucleic Acids Res* **32**,
867 D493-6 (2004).
868 84. Bolger, A.M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina
869 sequence data. *Bioinformatics* **30**, 2114-20 (2014).
870 85. Langmead, B. & Salzberg, S.L. Fast gapped-read alignment with Bowtie 2. *Nat Methods*
871 **9**, 357-9 (2012).
872 86. Love, M.I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion
873 for RNA-seq data with DESeq2. *Genome Biol* **15**, 550 (2014).
874 87. Team, R.C. R: A Language and Environment for Statistical Computing. (2019).

875 88. Pruitt, K.D., Tatusova, T. & Maglott, D.R. NCBI Reference Sequence (RefSeq): a
876 curated non-redundant sequence database of genomes, transcripts and proteins.
877 *Nucleic Acids Res* **33**, D501-4 (2005).
878 89. Cairns, J. *et al.* CHiCAGO: robust detection of DNA looping interactions in Capture Hi-C
879 data. *Genome Biol* **17**, 127 (2016).
880 90. Jung, I. *et al.* A compendium of promoter-centered long-range chromatin interactions in
881 the human genome. *Nat Genet* **51**, 1442-1449 (2019).
882 91. Kerimov, N. *et al.* eQTL Catalogue: a compendium of uniformly processed human gene
883 expression and splicing QTLs. *bioRxiv*, 2020.01.29.924266 (2020).
884 92. Alasoo, K. *et al.* Shared genetic effects on chromatin and gene expression indicate a
885 role for enhancer priming in immune response. *Nat Genet* **50**, 424-431 (2018).
886 93. Chen, L. *et al.* Genetic Drivers of Epigenetic and Transcriptional Variation in Human
887 Immune Cells. *Cell* **167**, 1398-1414 e24 (2016).
888 94. Gutierrez-Arcelus, M. *et al.* Passive and active DNA methylation and the interplay with
889 genetic variation in gene regulation. *Elife* **2**, e00523 (2013).
890 95. Lappalainen, T. *et al.* Transcriptome and genome sequencing uncovers functional
891 variation in humans. *Nature* **501**, 506-11 (2013).
892 96. Kilpinen, H. *et al.* Common genetic variation drives molecular heterogeneity in human
893 iPSCs. *Nature* **546**, 370-375 (2017).
894 97. Nedelec, Y. *et al.* Genetic Ancestry and Natural Selection Drive Population Differences
895 in Immune Responses to Pathogens. *Cell* **167**, 657-669 e21 (2016).
896 98. Quach, H. *et al.* Genetic Adaptation and Neandertal Admixture Shaped the Immune
897 System of Human Populations. *Cell* **167**, 643-656 e17 (2016).
898 99. Schwartzentruber, J. *et al.* Molecular and functional variation in iPSC-derived sensory
899 neurons. *Nat Genet* **50**, 54-61 (2018).
900 100. Buil, A. *et al.* Gene-gene and gene-environment interactions detected by transcriptome
901 sequence analysis in twins. *Nat Genet* **47**, 88-91 (2015).
902 101. van de Bunt, M. *et al.* Transcript Expression Data from Human Islets Links Regulatory
903 Signals from Genome-Wide Association Studies for Type 2 Diabetes and Glycemic
904 Traits to Their Downstream Effectors. *PLoS Genet* **11**, e1005694 (2015).
905 102. Schmiedel, B.J. *et al.* Impact of Genetic Polymorphisms on Human Immune Cell Gene
906 Expression. *Cell* **175**, 1701-1715 e16 (2018).
907 103. Jaffe, A.E. *et al.* Developmental and genetic regulation of the human cortex
908 transcriptome illuminate schizophrenia pathogenesis. *Nat Neurosci* **21**, 1117-1125
909 (2018).
910 104. Ng, B. *et al.* An xQTL map integrates the genetic architecture of the human brain's
911 transcriptome and epigenome. *Nat Neurosci* **20**, 1418-1426 (2017).
912 105. Lepik, K. *et al.* C-reactive protein upregulates the whole blood expression of CD59 - an
913 integrative analysis. *PLoS Comput Biol* **13**, e1005766 (2017).
914 106. Taylor, D.L. *et al.* Integrative analysis of gene expression, DNA methylation,
915 physiological traits, and genetic variation in human skeletal muscle. *Proc Natl Acad Sci*
916 *U S A* **116**, 10883-10888 (2019).
917 107. Fairfax, B.P. *et al.* Genetics of gene expression in primary immune cells identifies cell
918 type-specific master regulators and roles of HLA alleles. *Nat Genet* **44**, 502-10 (2012).
919 108. Fairfax, B.P. *et al.* Innate immune activity conditions the effect of regulatory variants
920 upon monocyte gene expression. *Science* **343**, 1246949 (2014).
921 109. Kasela, S. *et al.* Pathogenic implications for autoimmune mechanisms derived by
922 comparative eQTL analysis of CD4+ versus CD8+ T cells. *PLoS Genet* **13**, e1006643
923 (2017).
924 110. Naranbhai, V. *et al.* Genomic modulators of gene expression in human neutrophils. *Nat*
925 *Commun* **6**, 7545 (2015).

111. Degner, J.F. *et al.* DNase I sensitivity QTLs are a major determinant of human expression variation. *Nature* **482**, 390-4 (2012).
112. 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).
113. Raj, T. *et al.* Polarization of the effects of autoimmune and neurodegenerative risk alleles in leukocytes. *Science* **344**, 519-23 (2014).
114. Lee, M.N. *et al.* Common genetic variants modulate pathogen-sensing responses in human dendritic cells. *Science* **343**, 1246980 (2014).
115. Chen, E.Y. *et al.* Enrichr: interactive and collaborative HTML5 gene list enrichment analysis tool. *BMC Bioinformatics* **14**, 128 (2013).
116. Kuleshov, M.V. *et al.* Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. *Nucleic Acids Res* **44**, W90-7 (2016).
117. Barrett, T. *et al.* NCBI GEO: archive for functional genomics data sets--update. *Nucleic Acids Res* **41**, D991-5 (2013).
118. Andrews, S. FastQC: a quality control tool for high throughput sequence data. (2010).
119. 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).
120. Bentley, D.R. *et al.* Accurate whole human genome sequencing using reversible terminator chemistry. *Nature* **456**, 53-9 (2008).
121. Patwardhan, R.P. *et al.* High-resolution analysis of DNA regulatory elements by synthetic saturation mutagenesis. *Nat Biotechnol* **27**, 1173-5 (2009).
122. Vaughn, S.E. *et al.* Lupus risk variants in the PXK locus alter B-cell receptor internalization. *Front Genet* **5**, 450 (2014).
123. 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).
124. Amemiya, H.M., Kundaje, A. & Boyle, A.P. The ENCODE Blacklist: Identification of Problematic Regions of the Genome. *Sci Rep* **9**, 9354 (2019).
125. Lambert, S.A. *et al.* The Human Transcription Factors. *Cell* **172**, 650-665 (2018).
126. Stormo, G.D. Consensus patterns in DNA. *Methods Enzymol* **183**, 211-21 (1990).

957 **Figure Legends**

958 **Figure 1. Massively Parallel Reporter Assays Workflow.** See text for description.

959 **Figure 2. Regulatory activity of enhancer variants (enVars)**. a. Distribution of MPRA
960 regulatory activity. The normalized fold-change of MPRA activity relative to plasmid control (X-
961 axis) was calculated using DESeq2. Enhancer alleles (enAlleles) were identified as those alleles
962 with significant activity relative to control ($p_{adj} < 0.05$) and at least a 50% increase in activity (see
963 Methods). b. Enrichment of histone marks in GM12878 cell at enVars compared to non-enVars.
964 p-values were estimated using RELI (see Methods). Full RELI results are provided in
965 **Supplemental Data Set 9**. c. Enrichment of regulatory protein and transcription factor (TF)
966 binding at enVars compared to non-enVars. The top 15 TFs (based on RELI p-values) that overlap
967 at least 10% of enVars are shown. Full results are provided in **Supplemental Data Set 9**. d. TF
968 binding site motif enrichment for enVars compared to non-enVars. p-values were obtained from
969 HOMER using the full oligo sequences of enVars and non-enVars (see Methods). The top 15
970 enriched TF motif families are shown. Full results are provided in **Supplemental Data Set 10**.

971 **Figure 3. Regulatory activity of allelic enhancer variants (allelic enVars).** a. Identification of
972 allelic enVars. Genotype dependence (Y-axis) is defined as the normalized fold change of MPRA
973 activity between the non-reference and reference alleles (see Methods). MPRA enhancer activity
974 (X-axis) is presented as the maximum normalized fold-change of MPRA activity for any allele of
975 the variant. Allelic enVars were defined as variants with a significant difference in MPRA activity
976 ($p_{adj} < 0.05$) between any pair of alleles and at least a 25% change in activity difference (see
977 Methods). b. MPRA enhancer activity at the 27 risk loci with at least one allelic enVar. Bar plots
978 indicate the total number of variants at each locus. Variants with allelic enhancer activity (allelic
979 enVars) are shown in red. Variants lacking allelic enhancer activity are shown in grey.

980 **Figure 4. Lupus risk allele-dependent gene regulatory mechanisms at**
981 **the *C4A* and *SYNGR1* genomic loci.** a. and e. Normalized MPRA enhancer activity of each
982 experimental replicate for rs3101018 and rs26069235. b. and f. Expression trait quantitative loci
983 (eQTLs) revealing genotype-dependent expression of C4A and SNYGR1 for rs3101018 and
984 rs26069235 in EBV-transformed B cell lines (GTEx). c. and g. Genotype-dependent activity of
985 transcription factors, transcriptional regulators, and histone marks in EBV-transformed B cell
986 lines for rs3101018 and rs26069235. Results with MARIO ARS value >0.4 and consistent allelic
987 imbalance across ChIP-seq datasets are included (see Methods). The X-axis indicates the
988 preferred allele, along with a value indicating the strength of the allelic behavior, calculated as
989 one minus the ratio of the weak to strong read counts (e.g., 0.5 indicates the strong allele has
990 twice the reads of the weak allele). The median value is plotted when data from multiple cell
991 lines are available, with full results provided in **Supplemental Figure 4**. The numbers in
992 parentheses represent the number of ChIP-seq datasets with significant allelic activity (i.e. MARIO
993 ARS value > 0.4) out of the number of datasets where the given variant is inside a ChIP-seq peak
994 and is also heterozygous in the given cell line. *Variant overlapping* TFs are indicated in
995 black. *Variant adjacent* TFs are shown in green (see definition in **Figure 5a**). d. and h.
996 DNA binding motif logos are shown for the ATF/CREB/CREM family, and ELF1 in the context
997 of the DNA sequence surrounding rs3101018 and rs2069235, respectively. Tall nucleotides
998 above the X-axis indicate preferred DNA bases. Bases below the X-axis are disfavored.
999 **Figure 5. Identification of *variant overlapping* and *variant adjacent* TFs.** a. Model of *variant*
1000 *overlapping* and *variant adjacent* transcription factors (TFs). *Variant overlapping* TFs allelically
1001 bind on top of variants, while *variant adjacent* TFs allelically bind near variants. b. TF binding site
1002 location distribution for *variant overlapping* and *variant adjacent* TFs, relative to allelic enVars. c.
1003 TF motif families enriched for participating as *variant overlapping* TFs at allelic enVars. Motif
1004 disruption p-values were estimated by comparing the fraction of motif disruption events at allelic
1005 enVars to the fraction observed at non-allelic enVars (see Methods). d. TF motif families enriched

1006 for participating as *variant adjacent* TFs at allelic enVars. Motif enrichment p-values were
1007 estimated by comparing the fraction of predicted TF binding sites in allelic enVars to random
1008 expectation (see Methods). For both the *variant overlapping* and *variant adjacent* analyses, motif
1009 families are shown with p_{adj} <0.0001 and three or more allelic events at allelic enVar loci, or five
1010 or more predicted binding sites at allelic enVar loci, respectively.

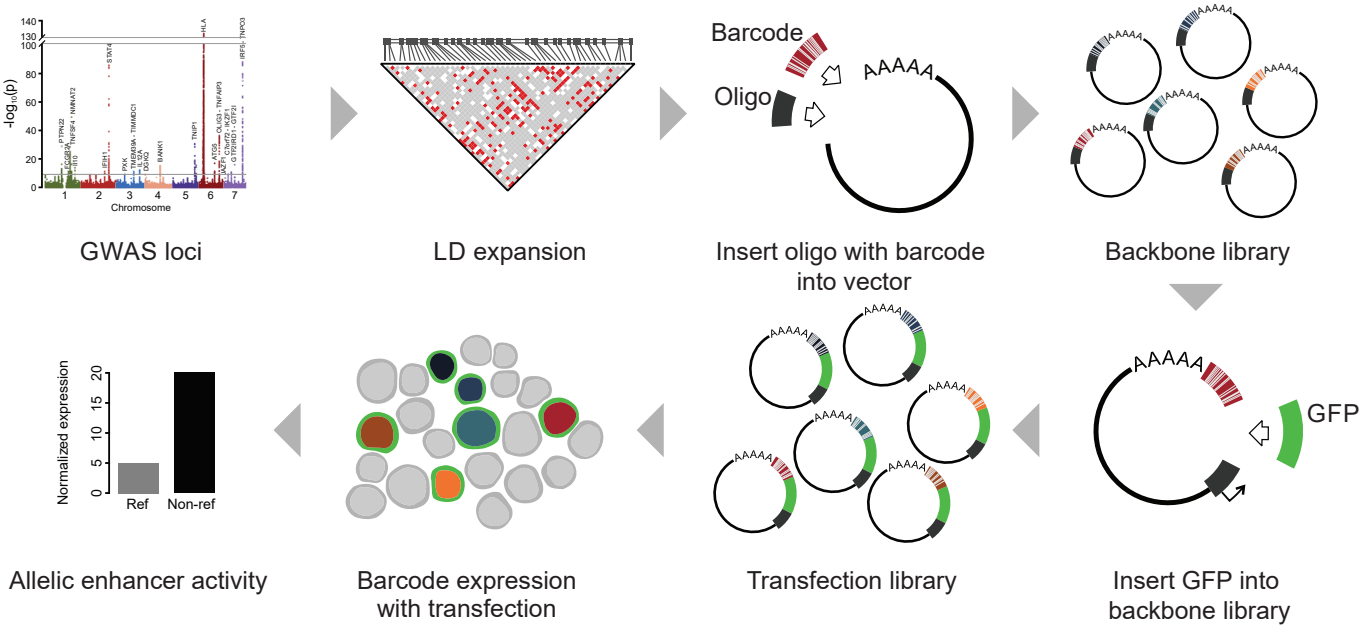
-log10(p)
Chromosome
GWAS loci
LD expansion
Barcode
Oligo
AAAAA
Insert oligo with barcode into vector
Backbone library
GFP
Insert GFP into backbone library
Transfection library
Barcode expression with transfection
Normalized expression
Ref
Non-ref
Allelic enhancer activity

Figure 1

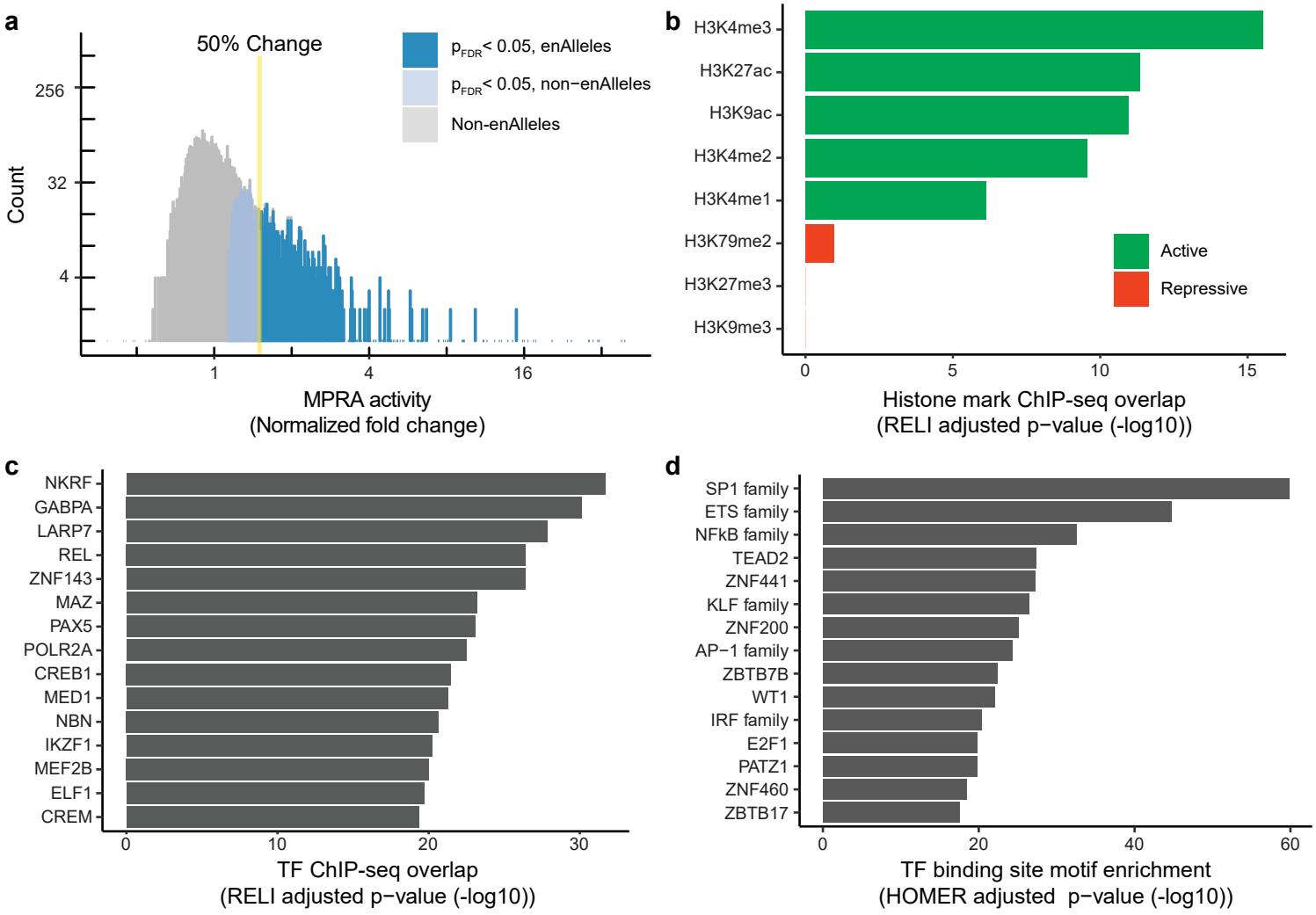
a
50% Change
p_{FDR}< 0.05, enAlleles
p_{FDR}< 0.05, non−enAlleles
Non-enAlleles
Count
256
32
4
1
4
16
MPRA activity
(Normalized fold change)
b
H3K4me3
H3K27ac
H3K9ac
H3K4me2
H3K4me1
H3K79me2
H3K27me3
H3K9me3
Active
Repressive
0
5
10
15
Histone mark ChIP-seq overlap
(RELI adjusted p−value (-log10))
c
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

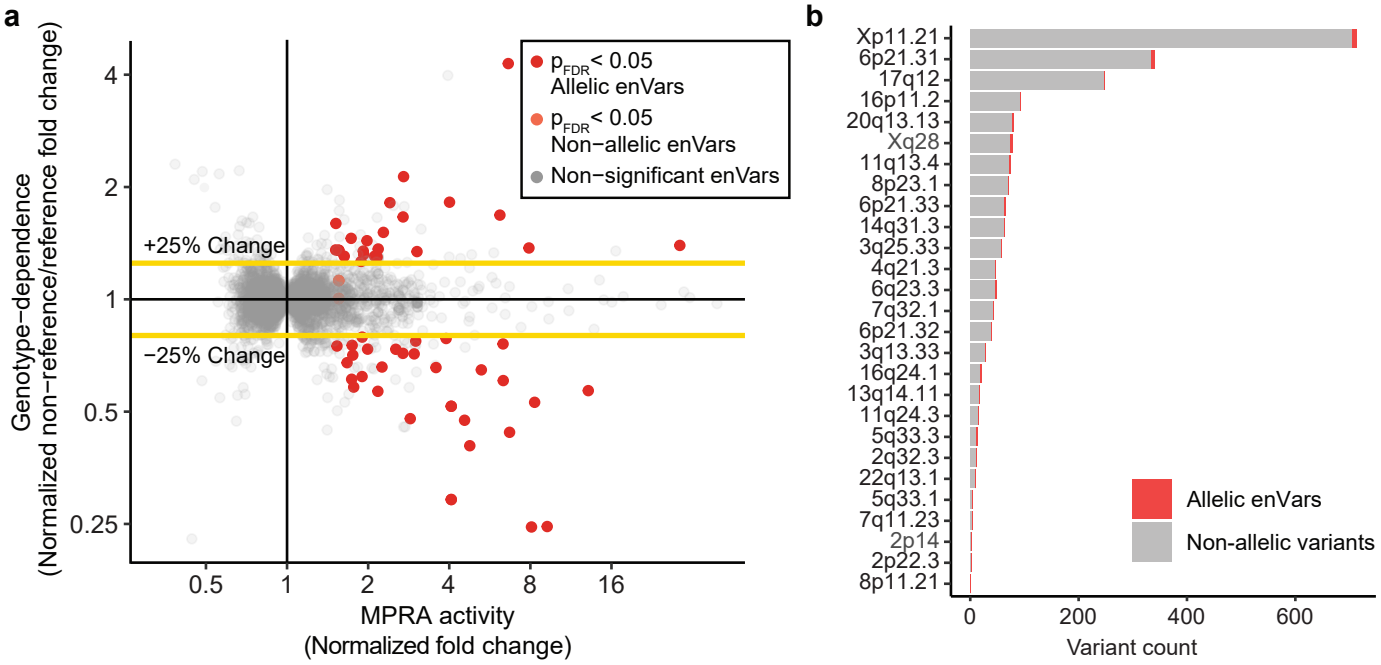
a
Genotype−dependence
(Normalized non−reference/reference fold change)
4
2
1
0.5
0.25
+25% Change
−25% Change
p_{FDR}< 0.05
Allelic enVars
p_{FDR}< 0.05
Non−allelic enVars
Non−significant enVars
0.5
1
2
4
8
16
MPRA activity
(Normalized fold change)
b
Xp11.21
6p21.31
17q12
16p11.2
20q13.13
Xq28
11q13.4
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

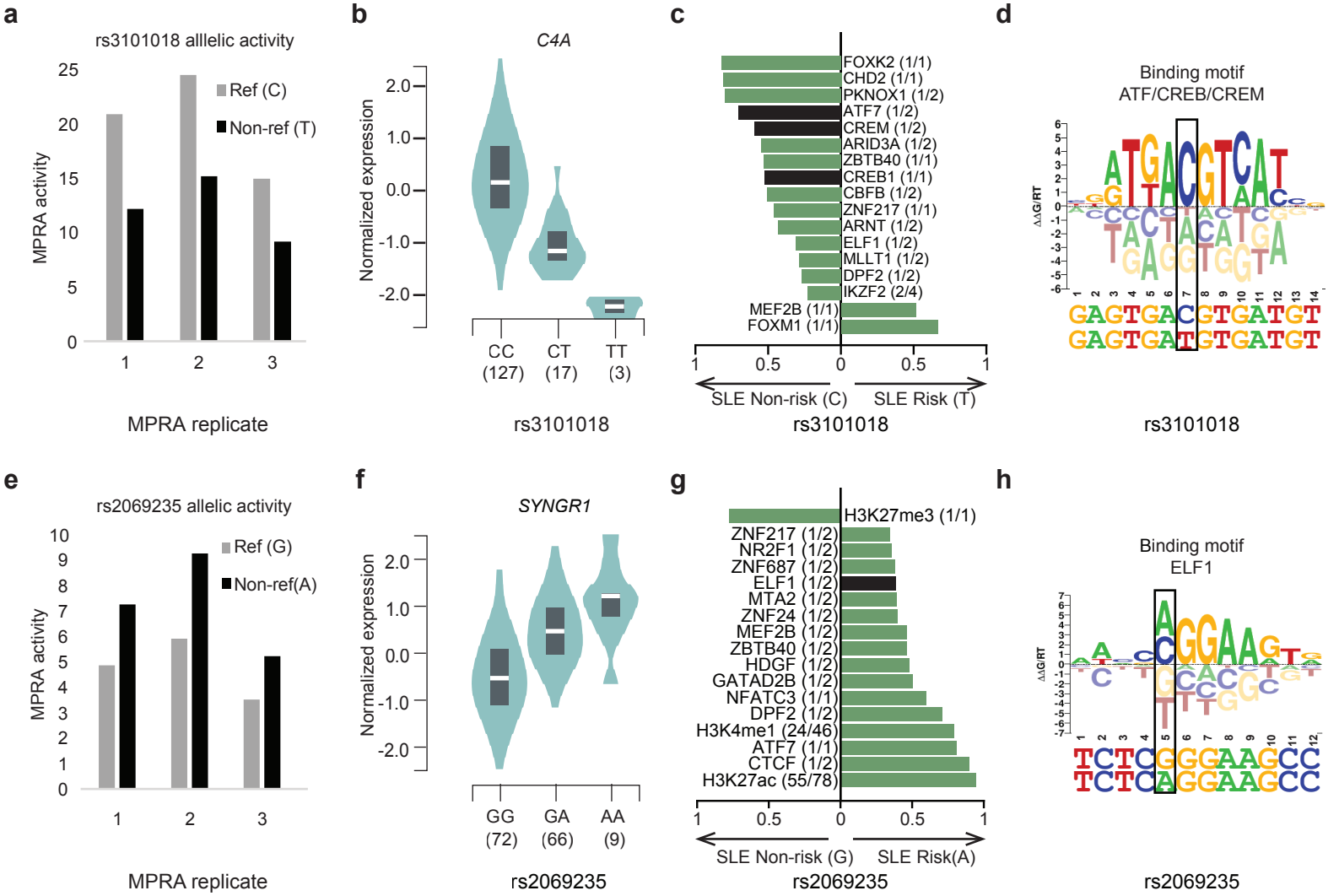
a
rs3101018 alllelic activity
MPRA activity
Ref (C)
Non-ref (T)
MPRA replicate
b
C4A
Normalized expression
CC (127)
CT (17)
TT (3)
rs3101018
c
FOXK2 (1/1)
CHD2 (1/1)
PKNOX1 (1/2)
ATF7 (1/2)
CREM (1/2)
ARID3A (1/2)
ZBTB40 (1/1)
CREB1 (1/1)
CBFB (1/2)
ZNF217 (1/1)
ARNT (1/2)
ELF1 (1/2)
MLLT1 (1/2)
DPF2 (1/2)
IKZF2 (2/4)
MEF2B (1/1)
FOXM1 (1/1)
SLE Non-risk (C)
SLE Risk (T)
rs3101018
d
Binding motif
ATF/CREB/CREM
ΔΔG/RT
rs3101018
e
rs2069235 allelic activity
MPRA activity
Ref (G)
Non-ref(A)
MPRA replicate
f
SYNGR1
Normalized expression
GG (72)
GA (66)
AA (9)
rs2069235
g
H3K27me3 (1/1)
ZNF217 (1/2)
NR2F1 (1/2)
ZNF687 (1/2)
ELF1 (1/2)
MTA2 (1/2)
ZNF24 (1/2)
MEF2B (1/2)
ZBTB40 (1/2)
HDGF (1/2)
GATAD2B (1/2)
NFATC3 (1/1)
DPF2 (1/2)
H3K4me1 (24/46)
ATF7 (1/1)
CTCF (1/2)
H3K27ac (55/78)
SLE Non-risk (G)
SLE Risk(A)
rs2069235
h
Binding motif
ELF1
ΔΔG/RT
rs2069235

Figure 4

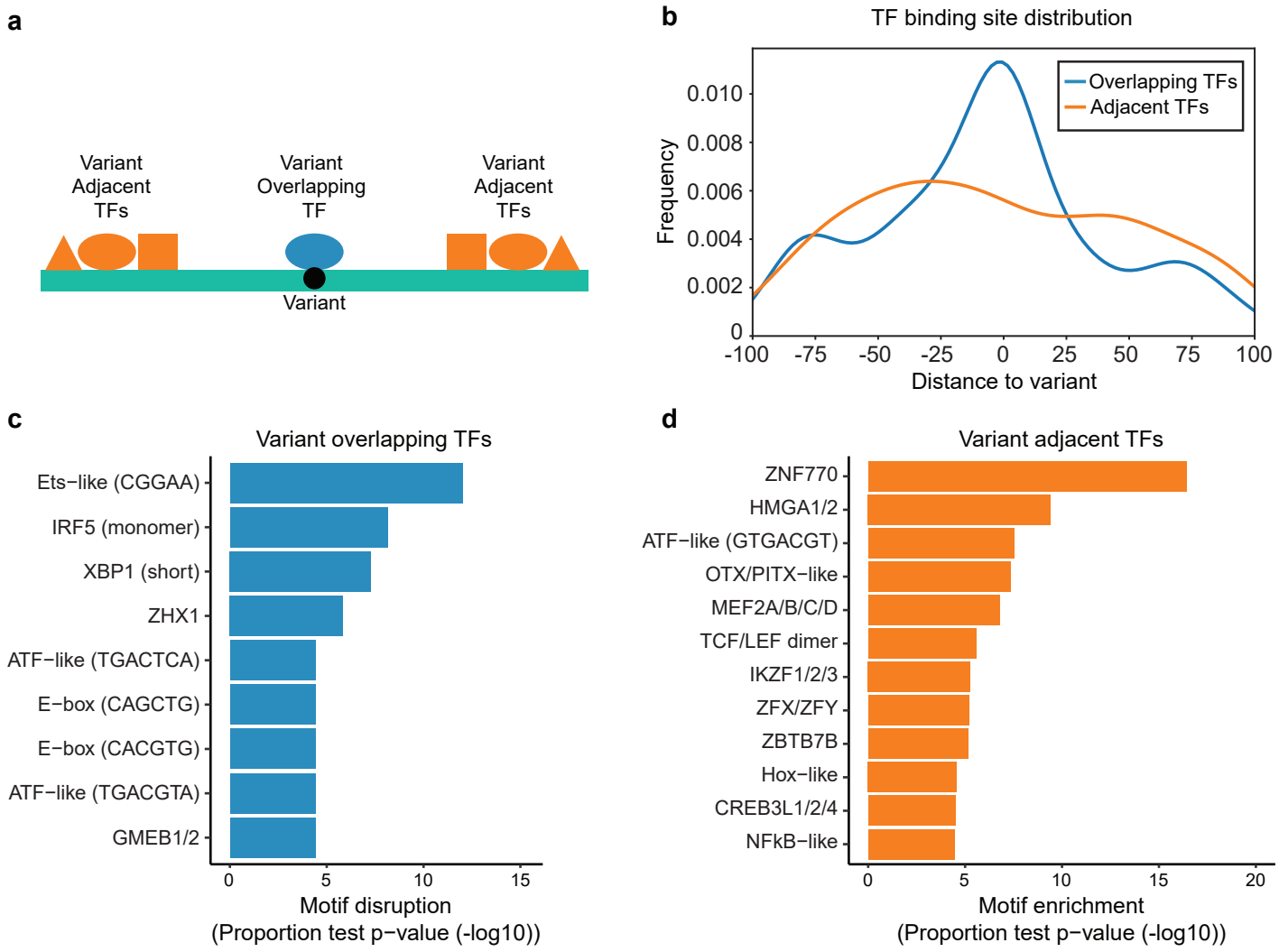
a
Variant Adjacent TFs
Variant Overlapping TF
Variant Adjacent TFs
Variant
b
TF binding site distribution
Overlapping TFs
Adjacent TFs
Frequency
0.010
0.008
0.006
0.004
0.002
0
-100
-75
-50
-25
0
25
50
75
100
Distance to variant
c
Variant overlapping TFs
Ets−like (CGGAA)
IRF5 (monomer)
XBP1 (short)
ZHX1
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